

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

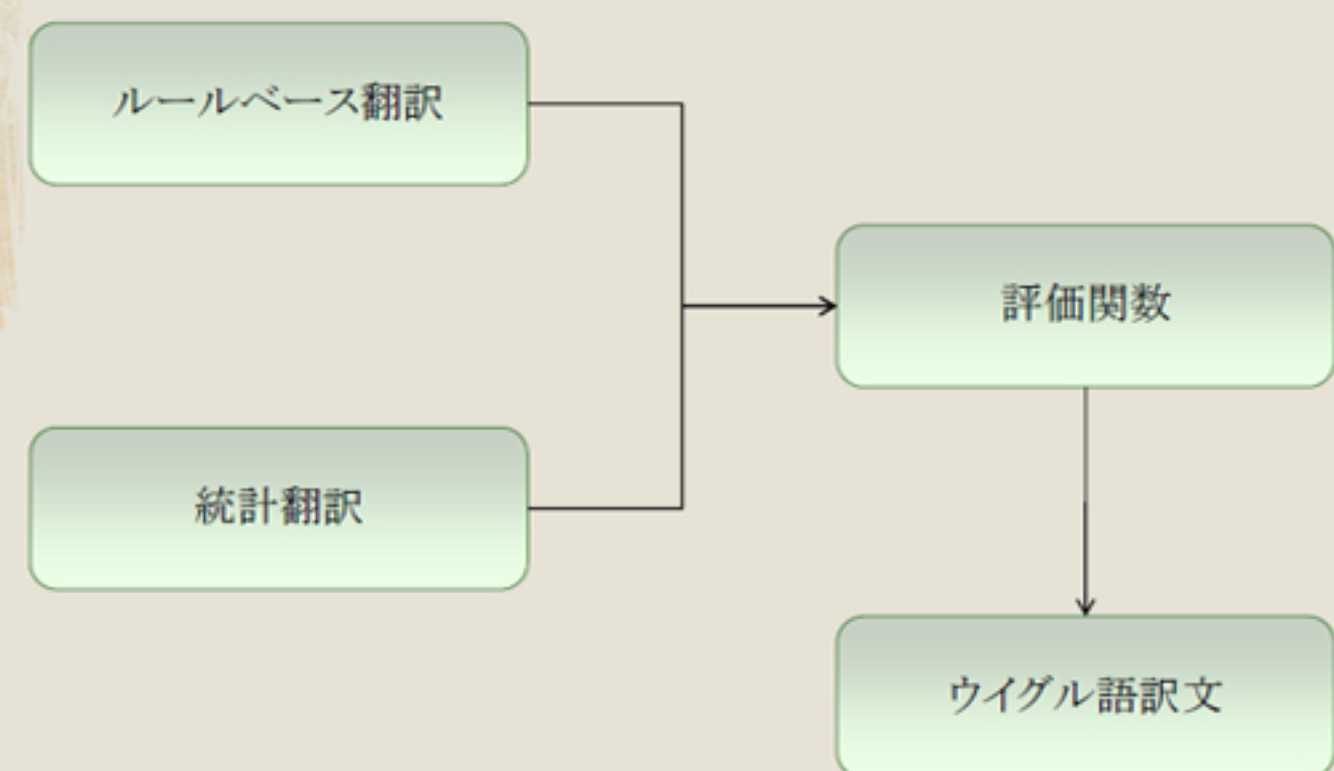
研究背景

- 日本語と英語、中国語など大きい言語の間機械翻訳が盛んに行われ、ウイグル語のような言語間での研究が少ない
- 最近統計翻訳が主流で、ウイグル語と他言語間での統計翻訳の研究が少ない状態
- 日本語ーウイグル語統計機械翻訳場合は学習コーパスが少ないこと
- ルールベース翻訳の場合各接辞の対応関係をルールで決めるのが大変、統計翻訳でそれらの対応関係を作らなくても済む

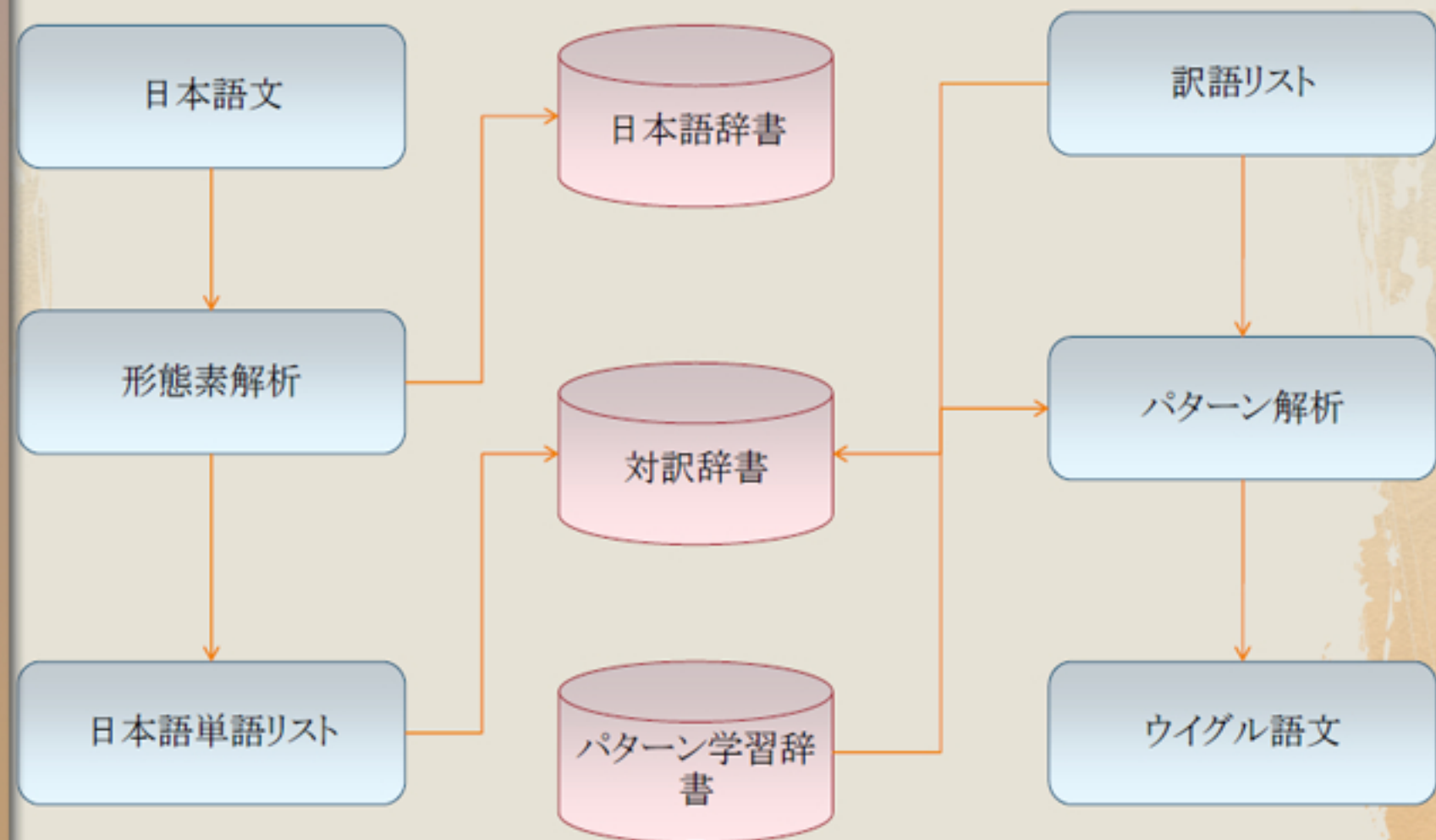
日本語とウイグル語の文法の相違点

	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	○	○

提案翻訳システムの流れ



提案翻訳システムの流れーMecabを拡張したルールベース翻訳システム



Mecab出力フォーマットの設定

鳥は遠い所から飛んで来ました

形態素	鳥	は	遠い	所	から	飛ん	で	来	まし	た
品詞1	名詞	助詞	形容詞	名詞	助詞	動詞	助詞	動詞	助動詞	助動詞
品詞2	一般	係助詞	自立	非自立	格助詞	自立	接統助詞	非自立		

対訳辞書から訳語を決定

- 日ーウ対訳辞書の作成:IPA辞書を基にし、日ーウ対訳辞書を作成する

VE.csv			
日本語	品詞	ウイグル語	品詞
言う	VE	diyish	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

対訳辞書から訳語を決定

- 日ーウ対訳辞書の作成:IPA辞書を基にし、日ーウ対訳辞書を作成する

鳥は遠い所から飛んで来ました

日本語	鳥	は	遠い	所	から	飛ん	で	来	まし	た
ウイグル語	qush	o	yeraq	jay	din	uchu	p	kel	o	di
File.csv	NU	POPC	ADJ	NU	CPOP	VE	COPOP	VE	AUXVE	AUXVE

ルールベースエンジンの作成

- ルールベースエンジンを作りことで以下の問題を解決することができる

単語の前後関係から、接続がくっつく語幹を決める
ウイグル語での母音と子音の弱化、脱落、差入などの問題の対応
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

ルールベースエンジンの作成

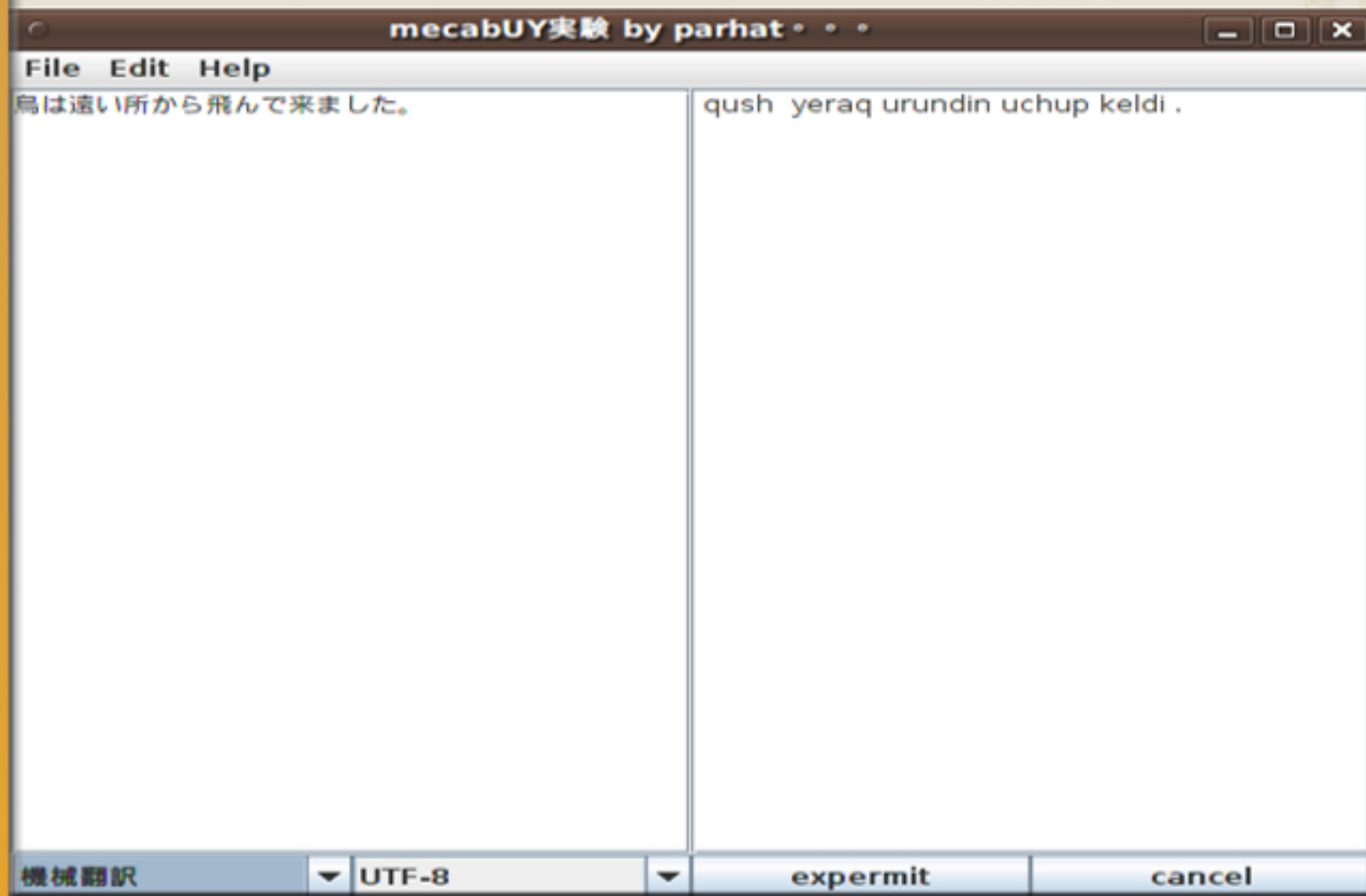
- ルールベースエンジンを作りことで以下の問題を解決することができる

Qush o yeraq jay din uchu p kel o di

NU	POPC	ADJ	NU	CPOP	VE	COPOP	VE	AUXVE	AUXVE

Qush yeraq jaydin uchup keldi

ルールベースエンジンの作成



言語モデルと翻訳モデルの設定

- 言語モデル:N-gramモデルの学習にSRILMを用いる、N-gram countを5に設定する
- 翻訳モデル:GIZA++を用いて単語アライメントの計算、句に基づく翻訳モデルの設定、Phrase tableの作成: graw-diag-final-and
- デコーダの設定:mosesを用いる。翻訳モデルの重みを4、言語モデルの重みを3、ほかはdefaultで設定する

翻訳モデルの設定

食べ物について昔から言われている常識もあります。
NULL (()) yimekliklerghe ((1)) nispten ((2)) burundin ((3 4 5 6)) dep ((7)) kilwatqan ((8 9)) kuzqarashlarmu ((10)) bar ((11 12)) . ((13))
Sentence pair (1570) source length 15 target length 22 alignment score : 3.559 22e-28
医学的に正しいものはほとんどないのですが、多くの人がまだ信じています。
NULL ((6 9)) dohturluq ((1)) ilmi ((2)) buyiche ((3)) toghra ((4)) ner siler ((5)) asasen ((7)) yoq ((8)) bolsimu ((10 11)) , ((12)) nurghun ((13 14)) kishler ((15 16)) yene (()) ishinip ((18 19)) yuruydu ((17 20 21)) . ((22))
Sentence pair (1571) source length 13 target length 19 alignment score : 8.634 15e-23
生活が変わっても、昔の常識をそのまま信じているのは変です。
NULL (()) turmush ((1)) uzgersimu ((2 3 4 5)) , ((6)) burunqi ((7 8)) k uzqarashlargha ((9 10 17 18)) shu (()) petti ((11)) ishinip ((12 13)) yur yuidighan ((14)) kishiler ((15 16)) hazirnu (()) bar (()) . ((19))
Sentence pair (1572) source length 18 target length 27 alignment score : 2.909 22e-38
時々その常識が正しいかどうか、どうしてその常識ができたか、考えてみたほうがいいでしょう。
NULL (()) bezi ((1)) chaghlarda (()) shu ((2 11)) kuzqarash ((3 12)) tog hrimu ((4 5 13 14)) qandaq ((6 7 8 16)) , ((9)) nimishqa ((10)) shundaq (()) kuz ((23)) qarashlar ((25 26)) bolghan ((15)) , ((17)) oylinip ((18)) baqqan ((20 21 22)) yahshi ((24)) he ((19)) . ((27))

学習データのまとめと実験結果

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

Baseline translation	12/50	24%
Replacement translation	19/50	38%
Rule based translation	22/50	44%

翻訳文事例

日本語	明日雨が降るそうです。	
Baseline translaion	ete yamghur kup 降る shundaq .	×
Replacement translation	ata yamghur * yaghi dighandek tur * .	△
Rule base 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 translaion	hiroshima din setiwalghan sowghatni keler hapte elip barimen .	○
Replacement translation	hiroshima de temlik * sowghat ni setiwal * di din , kelar hepte eliwalp p bar du .	△
Rule base translation	hiroshimade temlik sowghatni setiwal didin , kelar hepte eliwalp bar du .	△
日本語	彼はどこから来た？	
Baseline translaion	u yaqqa yerlerde yaghuz ?	×
Replacement translation	u * qeyer din kel di ?	△
Rule base translation	u qeyerdin keldi ?	○

まとめとこれからの課題

- 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人手によって評価を行った
- 評価手法としてMETEORとBLEUがあるが、今後の課題とする
- ルールベース翻訳:複雑な文に対して助詞と接辞の役割を決めるパターンを作ることが困難。
- 統計翻訳:対訳コーパスの量が少ない、単語アライメントの時助詞と接辞のずれが大きい、翻訳精度が低い
- 単一の文の評価を決めること
- 対訳コーパスの拡張とパターンの種類を増加
- 学習データを作成する際にウイグル語に関しても形態素ごとに分割すること。