MORSE: Morphological analysis using a sequence decoder

Ekin Akyürek, Erenay Dayanık, Deniz Yuret
Koç University Artificial Intelligence Lab

https://github.com/ai-ku/Morse.jl
https://github.com/ai-ku/TrMor2018

Overview

- Problem
- Model
- Experiments

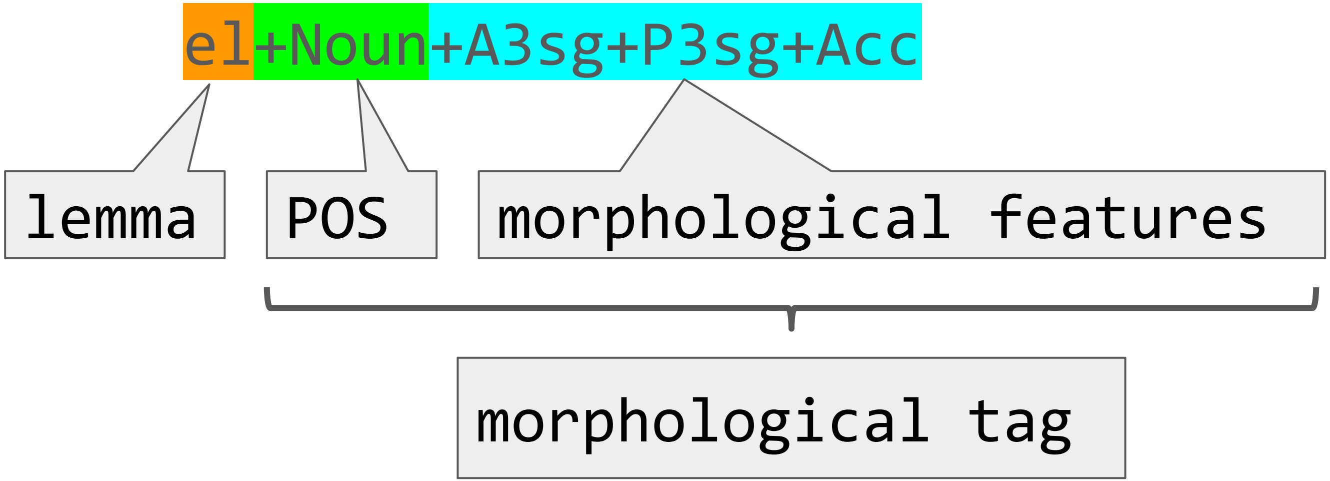
Problem

Goal: correct morphological analysis

Sonra	sonra+Adverb
gülerek	gül+Verb+Pos^DB+Adverb+ByDoingSo
elini	el+Noun+A3sg+P3sg+Acc
kardeşinin	kardeş+Noun+A3sg+P3sg+Gen
omzuna	omuz+Noun+A3sg+P3sg+Dat
koydu	koy+Verb+Pos+Past+A3sg

Components of morphological analysis (1/2)

“elini”



el+Noun+A3sg+P3sg+Acc

lemma

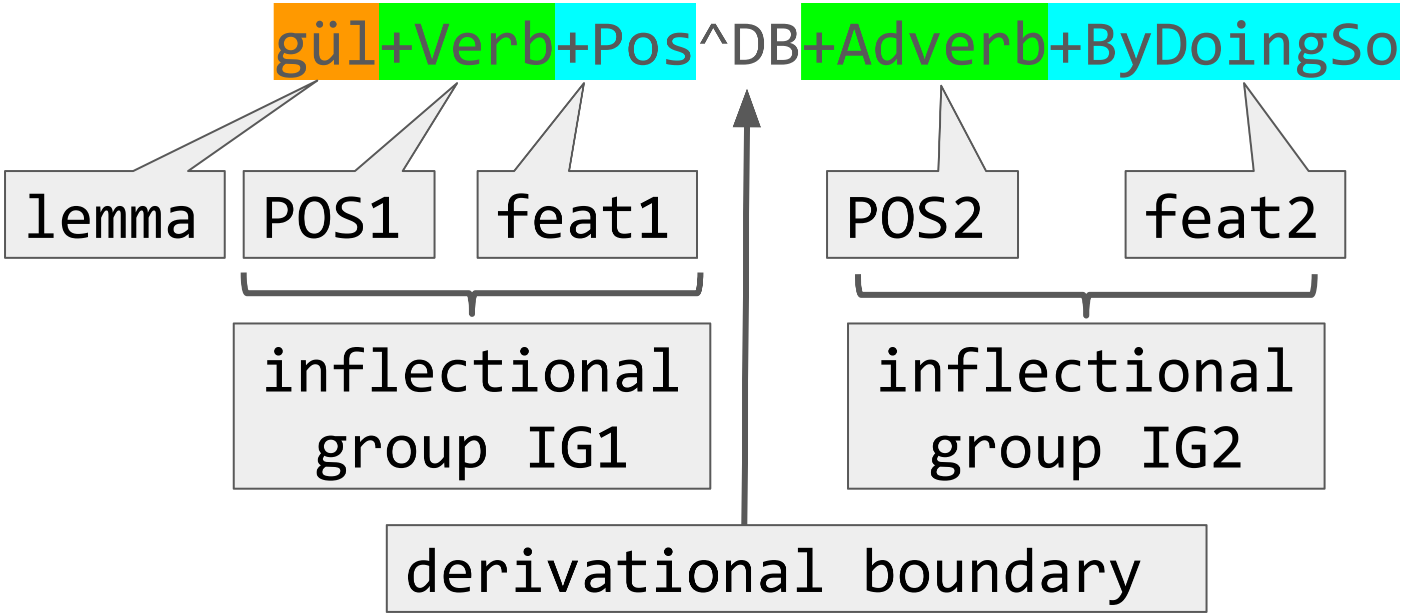
POS

morphological features

morphological tag

Components of morphological analysis (2/2)

“gülerek”



gül+Verb+Pos^DB+Adverb+ByDoingSo

lemma

POS1

feat1

POS2

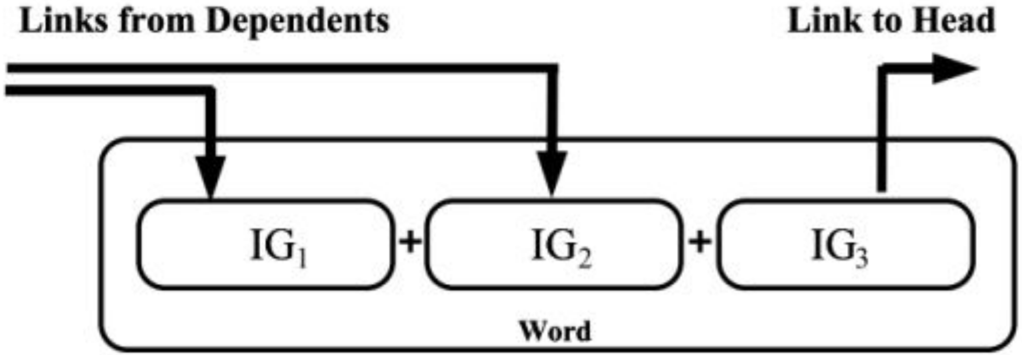
feat2

inflectional group IG1

inflectional group IG2

derivational boundary

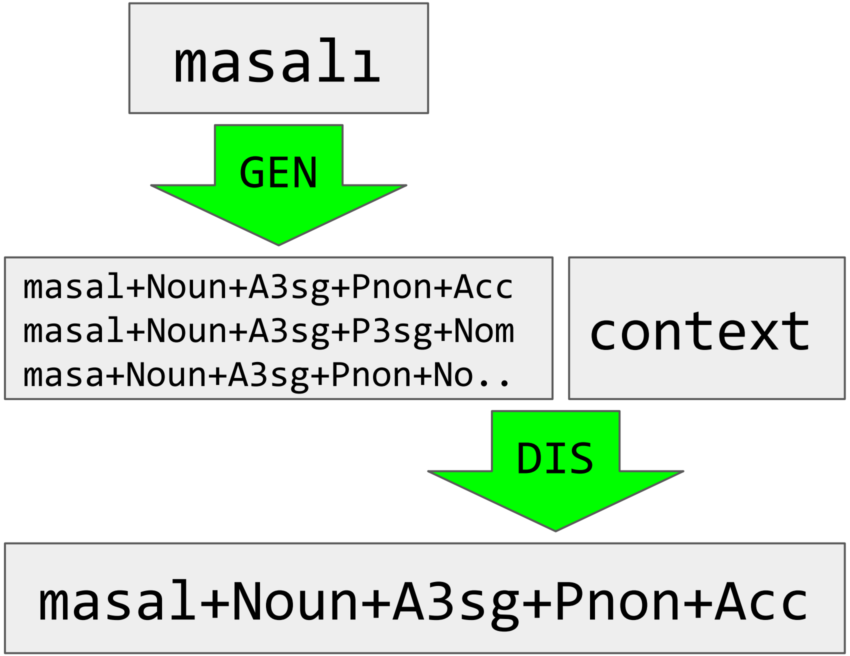
Multiple inflectional groups may have independent syntactic relationships



mavi masalı oda

Previous approaches (Morse does it all)

Generate & Disambiguate



masalı

GEN

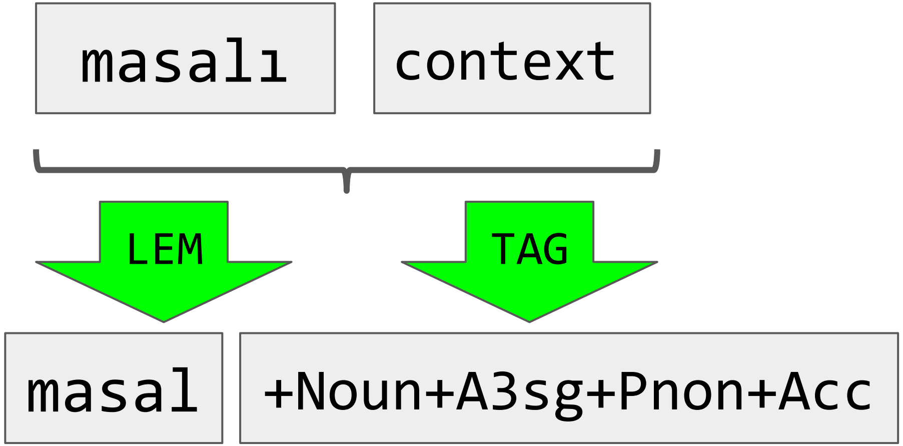
masal+Noun+A3sg+Pnon+Acc
masal+Noun+A3sg+P3sg+Nom
masa+Noun+A3sg+Pnon+No..

context

DIS

masal+Noun+A3sg+Pnon+Acc

Lemmatize & Tag



masalı

context

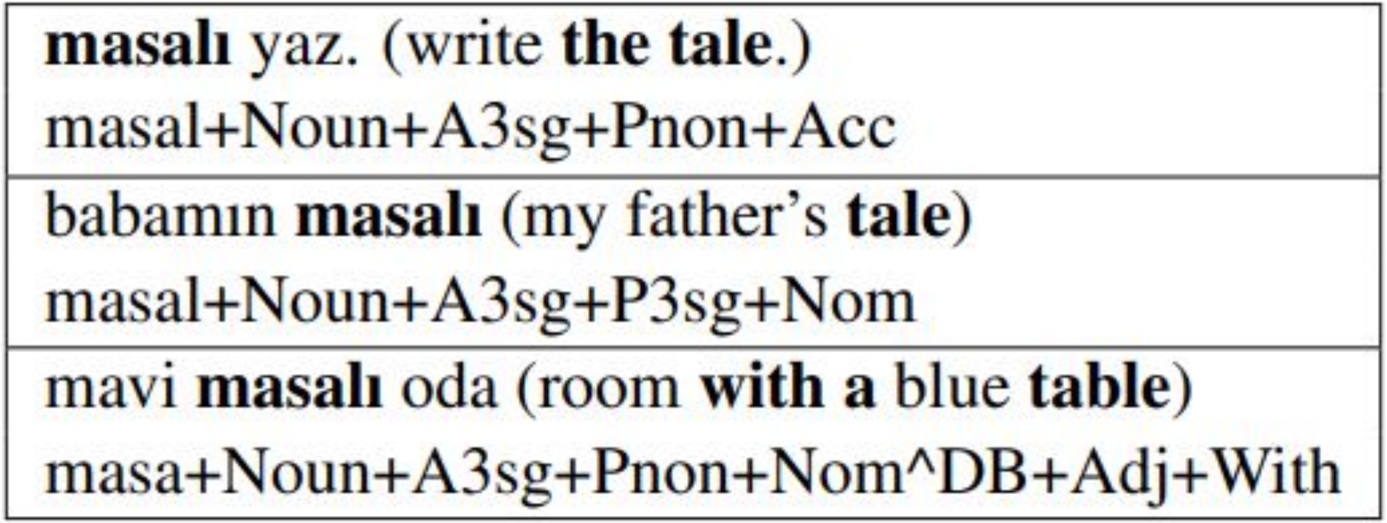
LEM

TAG

masal

+Noun+A3sg+Pnon+Acc

Challenge: morphological ambiguity



masalı yaz. (write **the tale**.) masal+Noun+A3sg+Pnon+Acc
babamın **masalı** (my father’s **tale**) masal+Noun+A3sg+P3sg+Nom
mavi **masalı** oda (room **with a** blue **table**) masa+Noun+A3sg+Pnon+Nom^DB+Adj+With

Three inputs that determine morphology

1. Word orthography
2. Semantic context
3. Syntactic context

Look at word orthography (1/3)

Sonra	sonra+Adverb
gülerek	gül+Verb+Pos^DB+Adverb+ByDoingSo
elini	el+Noun+A3sg+P3sg+Acc
kardeşinin	kardeş+Noun+A3sg+P3sg+Gen
omzuna	omuz+Noun+A3sg+P3sg+Dat
koydu	koy+Verb+Pos+Past+A3sg

Look at semantic context (2/3)

Sonra	sonra+Adverb
gülerek	gül+Verb+Pos^DB+Adverb+ByDoingSo
elini	el+Noun+A3sg+P3sg+Acc
kardeşinin	kardeş+Noun+A3sg+P3sg+Gen
omzuna	omuz+Noun+A3sg+P3sg+Dat
koydu	koy+Verb+Pos+Past+A3sg

Look at syntactic context (3/3)

Sonra	sonra+Adverb
gülerek	gül+Verb+Pos^DB+Adverb+ByDoingSo
elini	el+Noun+A3sg+P3sg+Acc
kardeşinin	kardeş+Noun+A3sg+P3sg+Gen
omzuna	omuz+Noun+A3sg+P3sg+Dat
koydu	koy+Verb+Pos+Past+A3sg

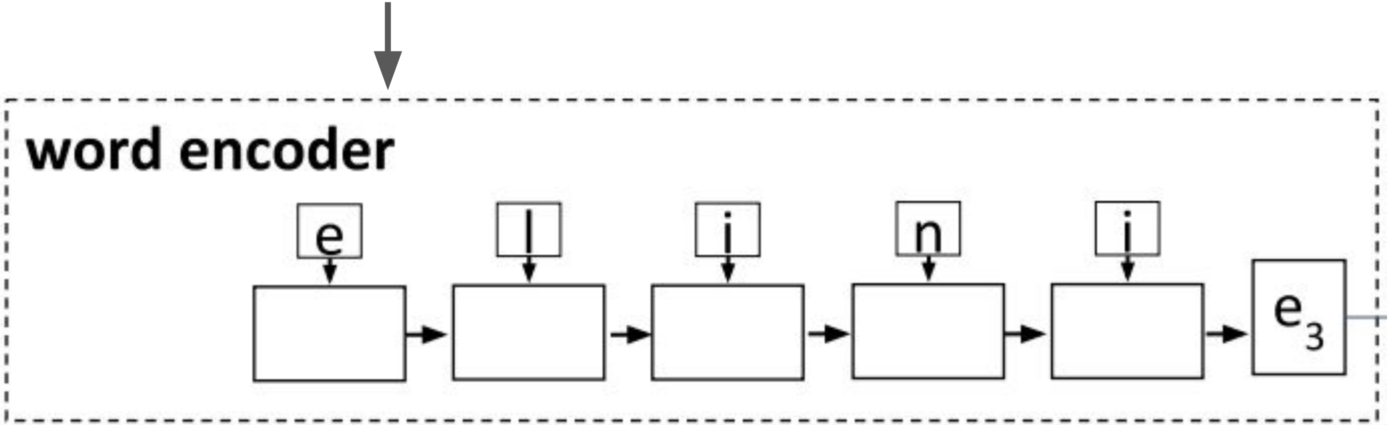
Model

The MORSE Model

- INPUT: Use RNNs to turn variable length sequences to fixed size feature vectors:
    - word encoder
    - context encoder
    - output encoder
- OUTPUT: Use a sequence decoder to produce lemma+tag one character/feature at a time.

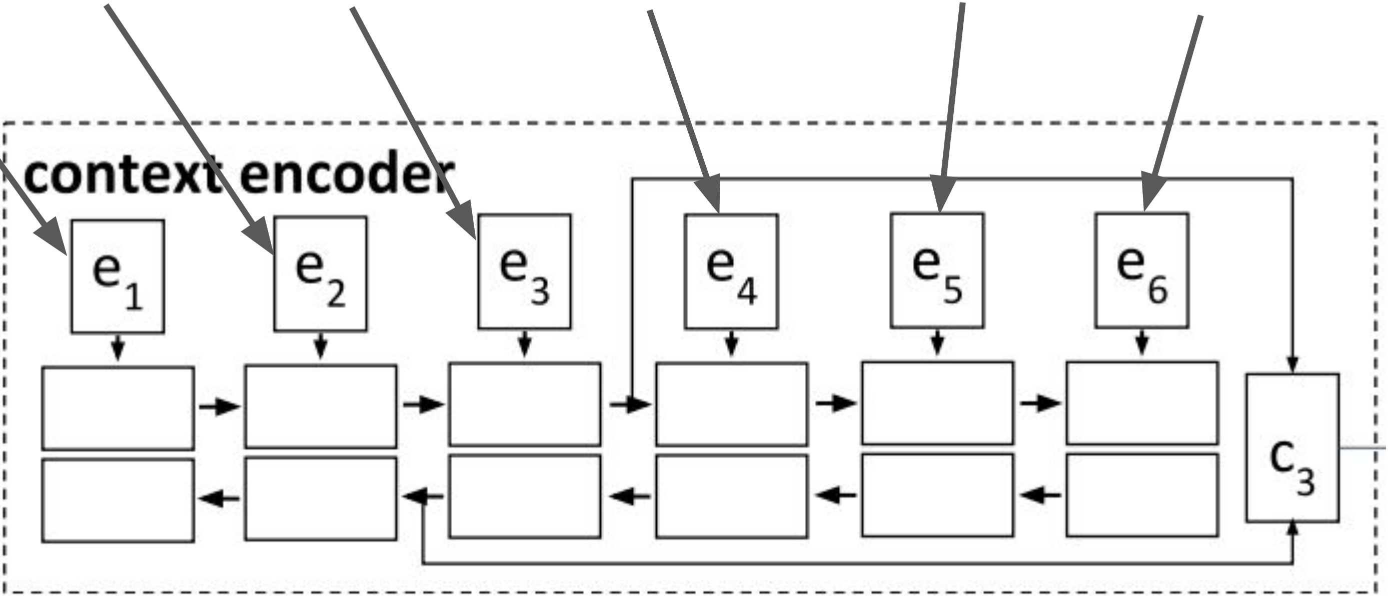
Encode words using RNN on characters (1/3)

Sonra gülerek elini kardeşinin omzuna koydu



Encode context using biRNN on words (2/3)

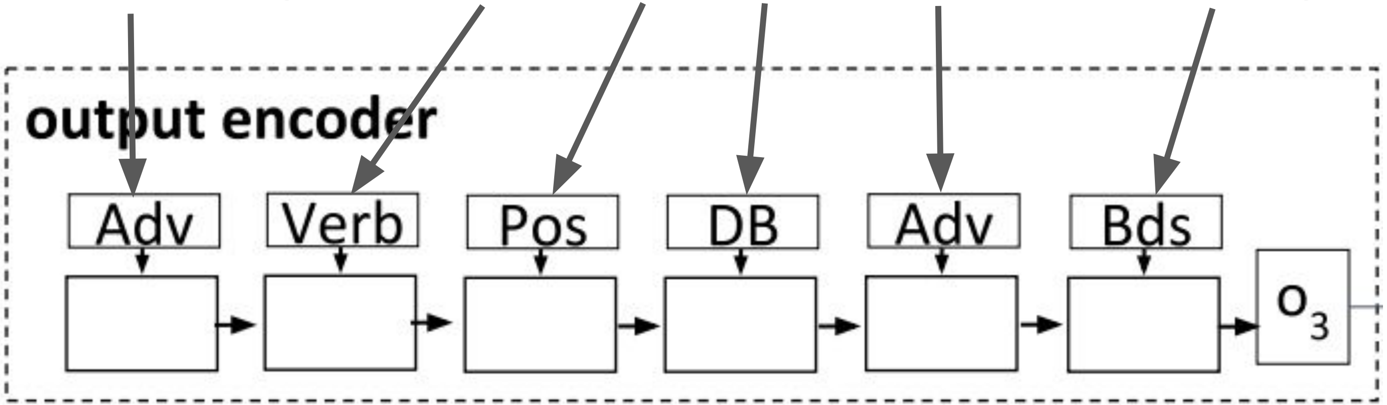
Sonra gülerek elini kardeşinin omzuna koydu



Encode output using RNN on prev tags (3/3)

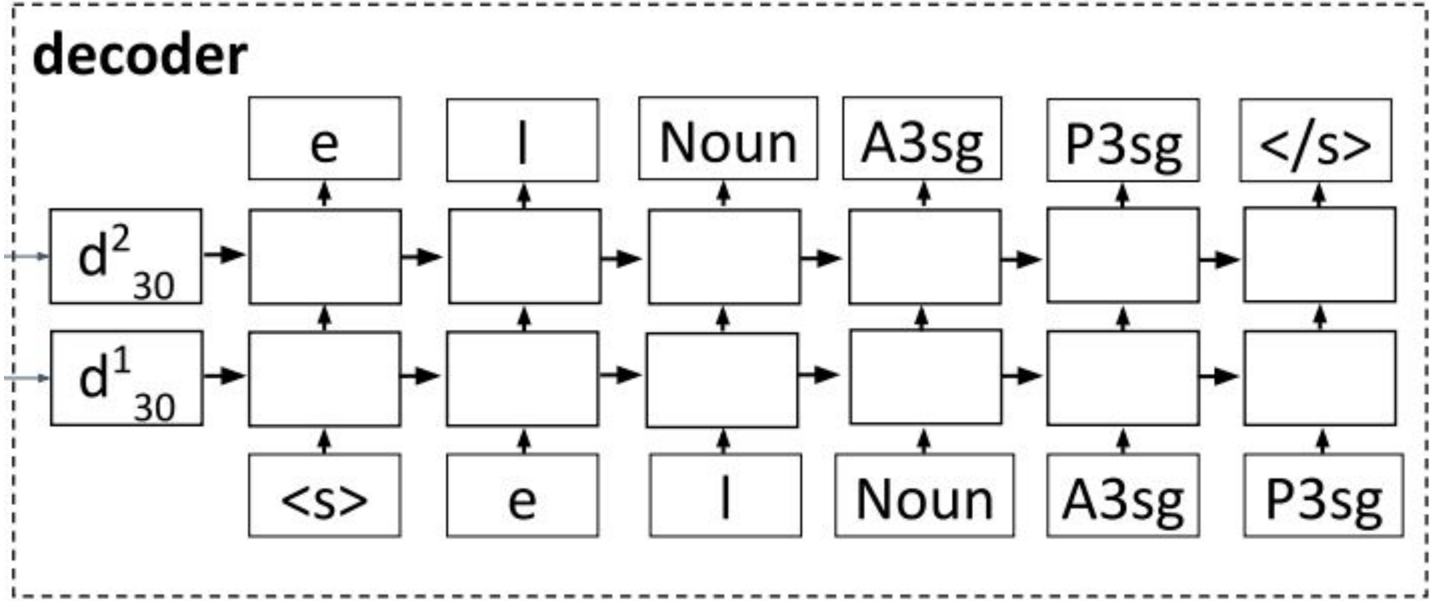
Sonra gülerek elini kardeşinin omzuna koydu

sonra+Adverb gül+Verb+Pos^DB+Adverb+ByDoingSo



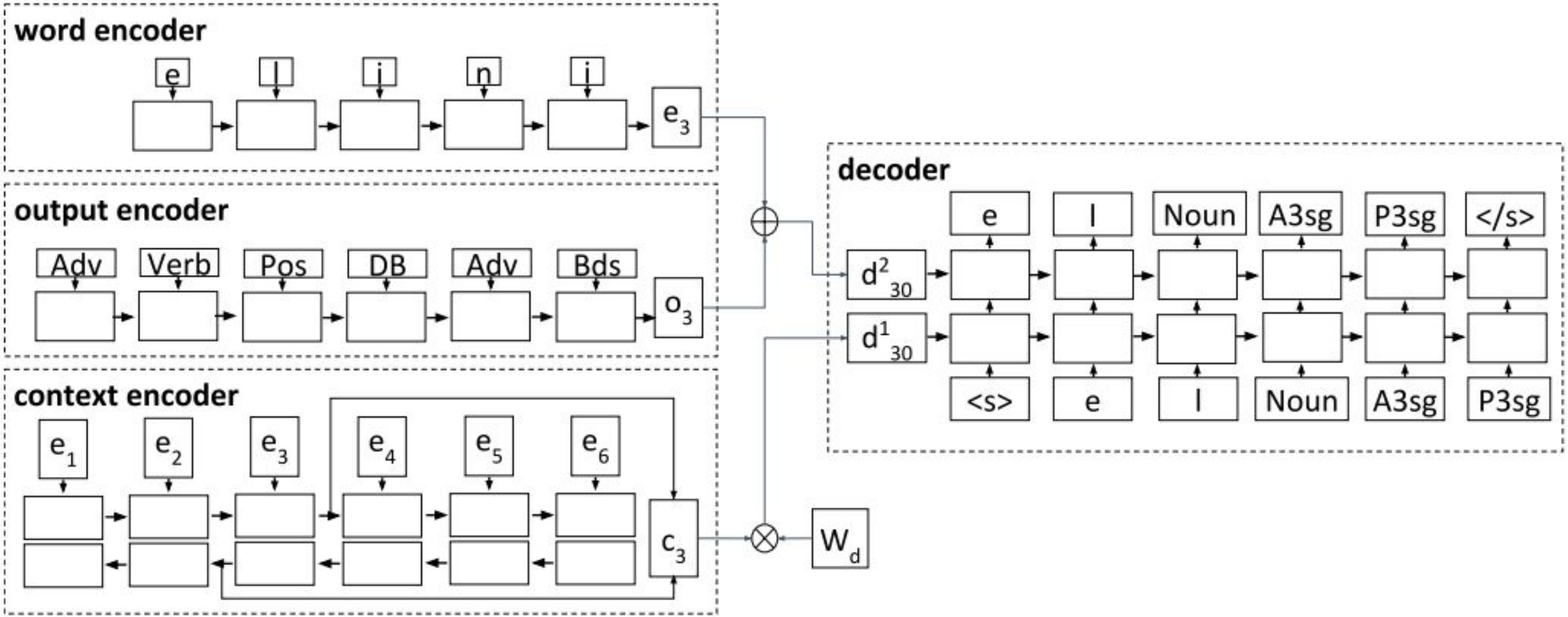
Produce one character/tag at a time

Sonra gülerek elini kardeşinin omzuna koydu



The whole architecture of Morse

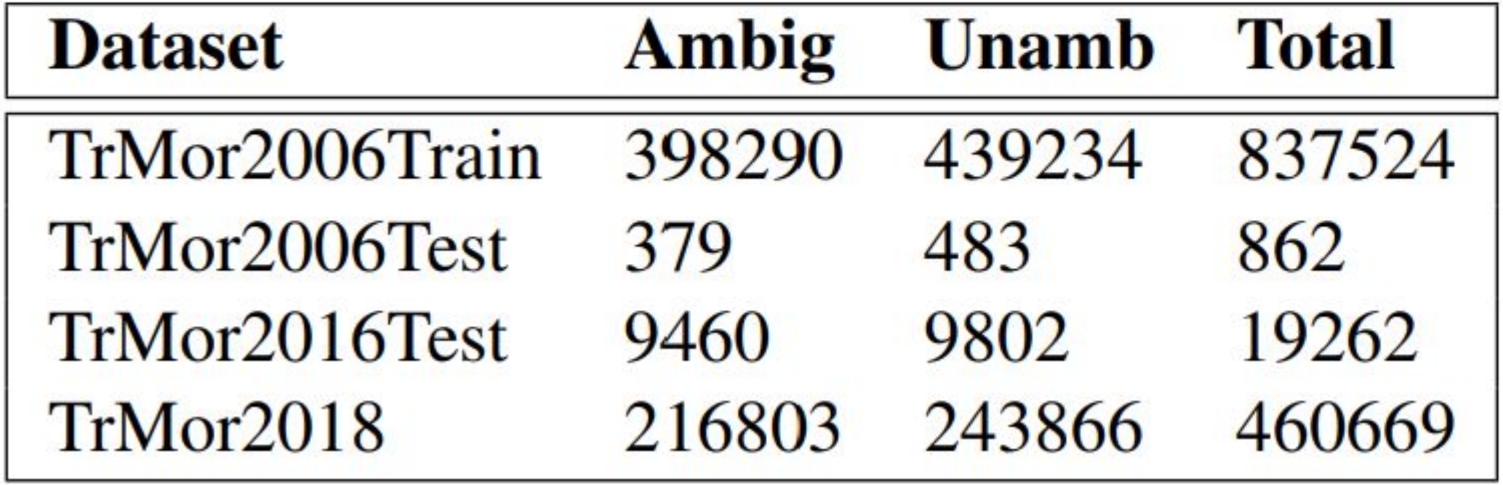
Sonra gülerek elini kardeşinin omzuna koydu



Experiments

Turkish datasets

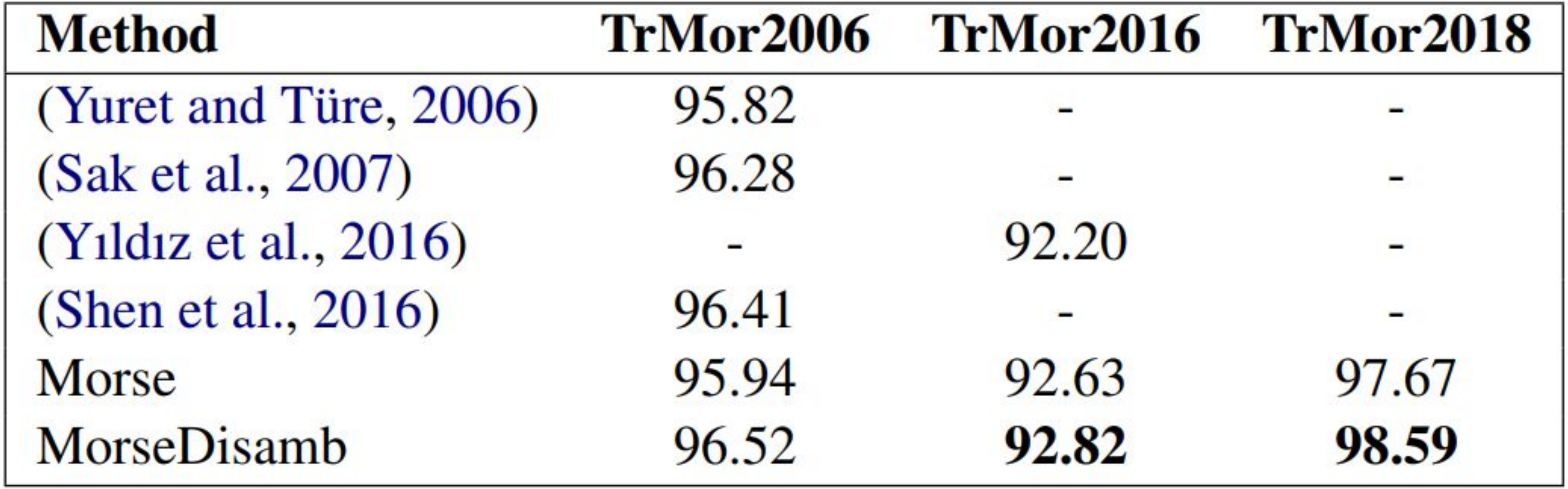
(TrMor2018 introduced in this work)



Dataset	Ambig	Unamb	Total
TrMor2006Train	398290	439234	837524
TrMor2006Test	379	483	862
TrMor2016Test	9460	9802	19262
TrMor2018	216803	243866	460669

Turkish experiments

(significant improvements in TrMor2016,TrMor2018)

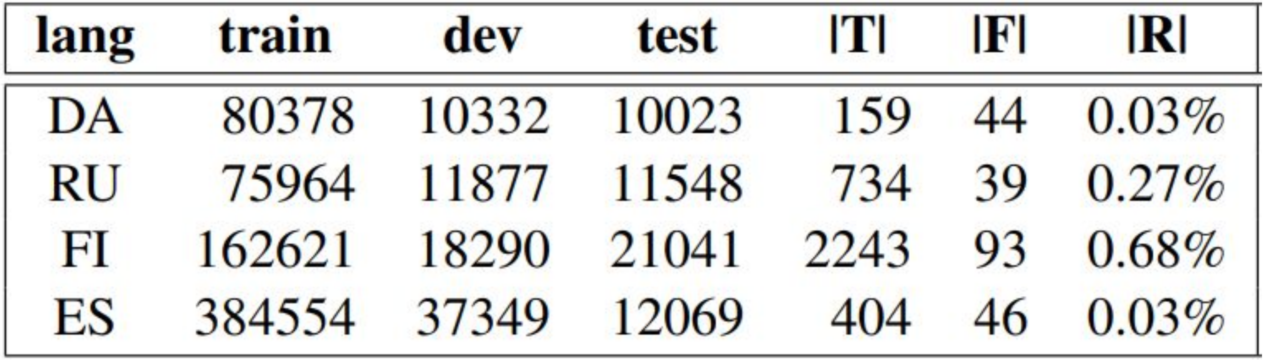


Method	**TrMor2006**	**TrMor2016**	**TrMor2018**
(Yuret and Türe, 2006)	95.82	-	-
(Sak et al., 2007)	96.28	-	-
(Yıldız et al., 2016)	-	92.20	-
(Shen et al., 2016)	96.41	-	-
Morse	95.94	92.63	97.67
MorseDisamb	96.52	**92.82**	**98.59**

MorseDisamb uses Morse to score output of a morphological generator for fair comparison.

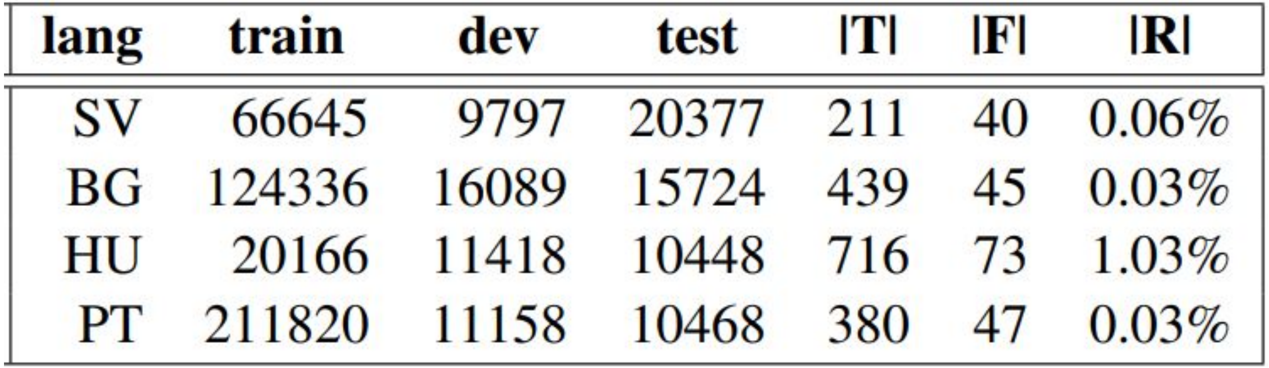
Multilingual datasets (UD 2.1)

High resource:



lang	train	dev	test	|T|	|F|	|R|
DA	80378	10332	10023	159	44	0.03%
RU	75964	11877	11548	734	39	0.27%
FI	162621	18290	21041	2243	93	0.68%
ES	384554	37349	12069	404	46	0.03%

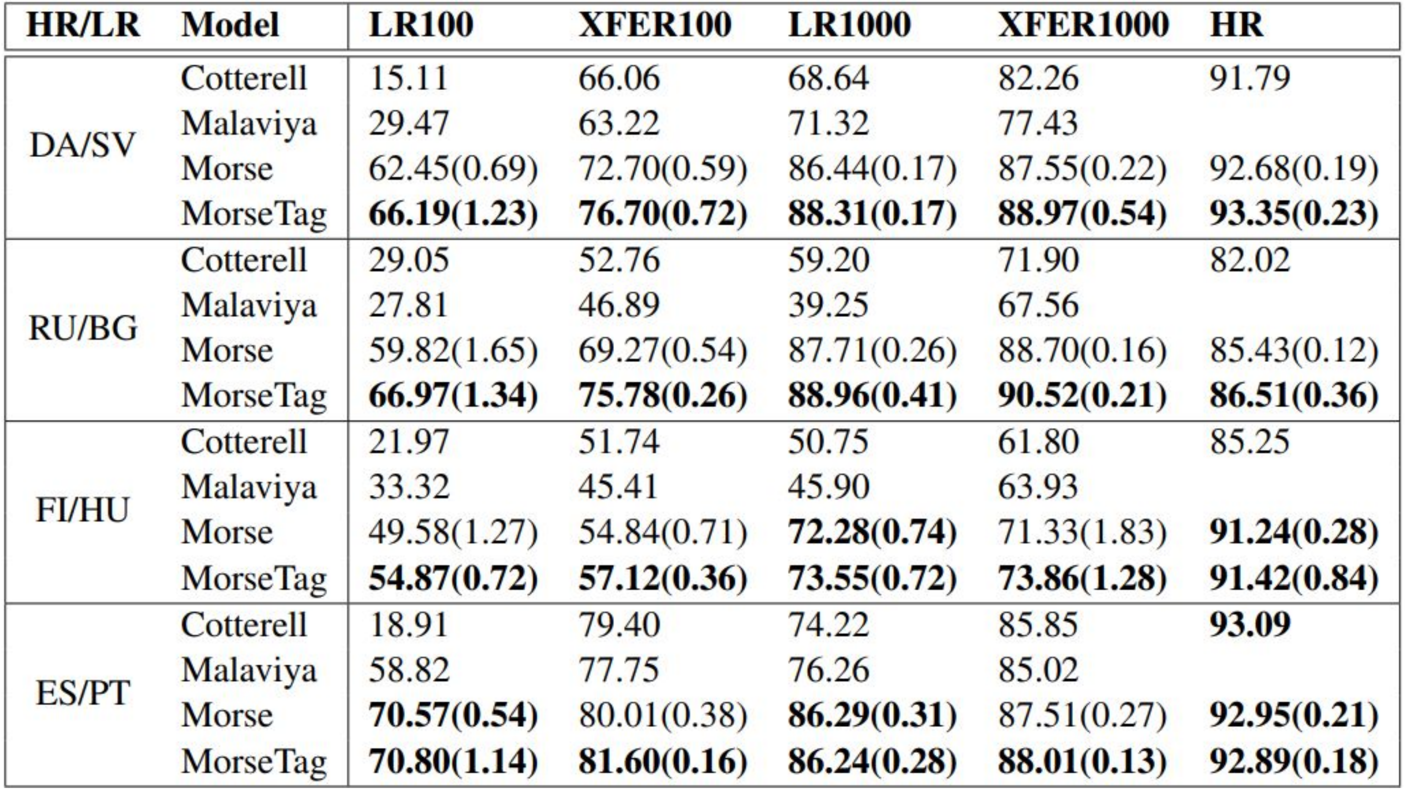
Low resource:



lang	train	dev	test	|T|	|F|	|R|
SV	66645	9797	20377	211	40	0.06%
BG	124336	16089	15724	439	45	0.03%
HU	20166	11418	10448	716	73	1.03%
PT	211820	11158	10468	380	47	0.03%

|T|: tags
|F|: features
|R|: unseen tag%

Multilingual experiments

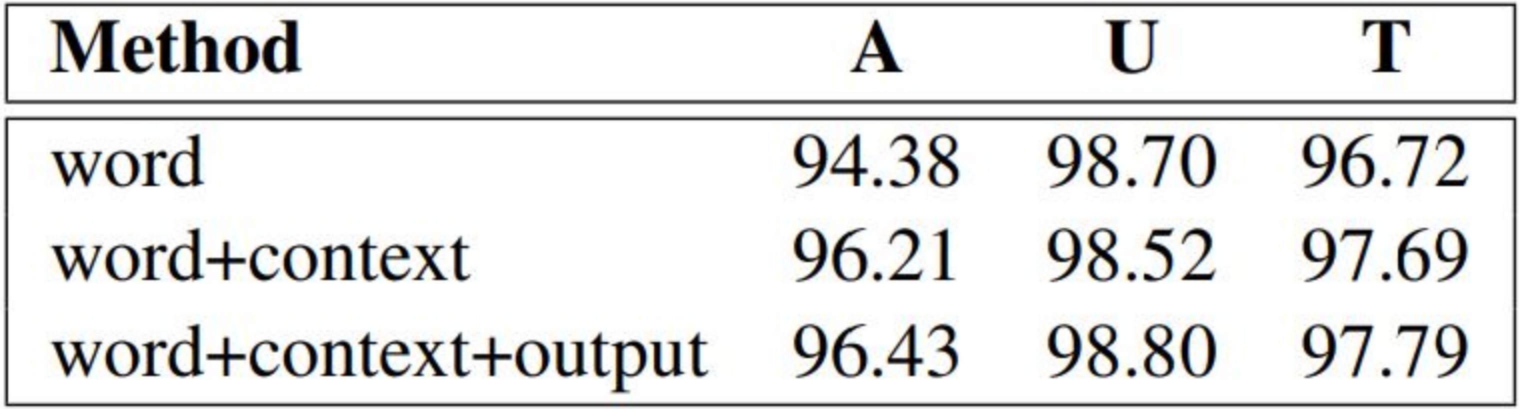


HR/LR	Model	LR100	XFER100	LR1000	XFER1000	HR
DA/SV	Cotterell	15.11	66.06	68.64	82.26	91.79
	Malaviya	29.47	63.22	71.32	77.43	
	Morse	62.45(0.69)	72.70(0.59)	86.44(0.17)	87.55(0.22)	92.68(0.19)
	MorseTag	**66.19(1.23)**	**76.70(0.72)**	**88.31(0.17)**	**88.97(0.54)**	**93.35(0.23)**
RU/BG	Cotterell	29.05	52.76	59.20	71.90	82.02
	Malaviya	27.81	46.89	39.25	67.56	
	Morse	59.82(1.65)	69.27(0.54)	87.71(0.26)	88.70(0.16)	85.43(0.12)
	MorseTag	**66.97(1.34)**	**75.78(0.26)**	**88.96(0.41)**	**90.52(0.21)**	**86.51(0.36)**
FI/HU	Cotterell	21.97	51.74	50.75	61.80	85.25
	Malaviya	33.32	45.41	45.90	63.93	
	Morse	49.58(1.27)	54.84(0.71)	**72.28(0.74)**	71.33(1.83)	**91.24(0.28)**
	MorseTag	**54.87(0.72)**	**57.12(0.36)**	**73.55(0.72)**	**73.86(1.28)**	**91.42(0.84)**
ES/PT	Cotterell	18.91	79.40	74.22	85.85	**93.09**
	Malaviya	58.82	77.75	76.26	85.02	
	Morse	**70.57(0.54)**	80.01(0.38)	**86.29(0.31)**	87.51(0.27)	**92.95(0.21)**
	MorseTag	**70.80(1.14)**	**81.60(0.16)**	**86.24(0.28)**	**88.01(0.13)**	**92.89(0.18)**

MorseTag uses Morse to only generate the morphological tag for fair comparison.

Ablation analysis on TrMor2018

(both context and output encoders help)

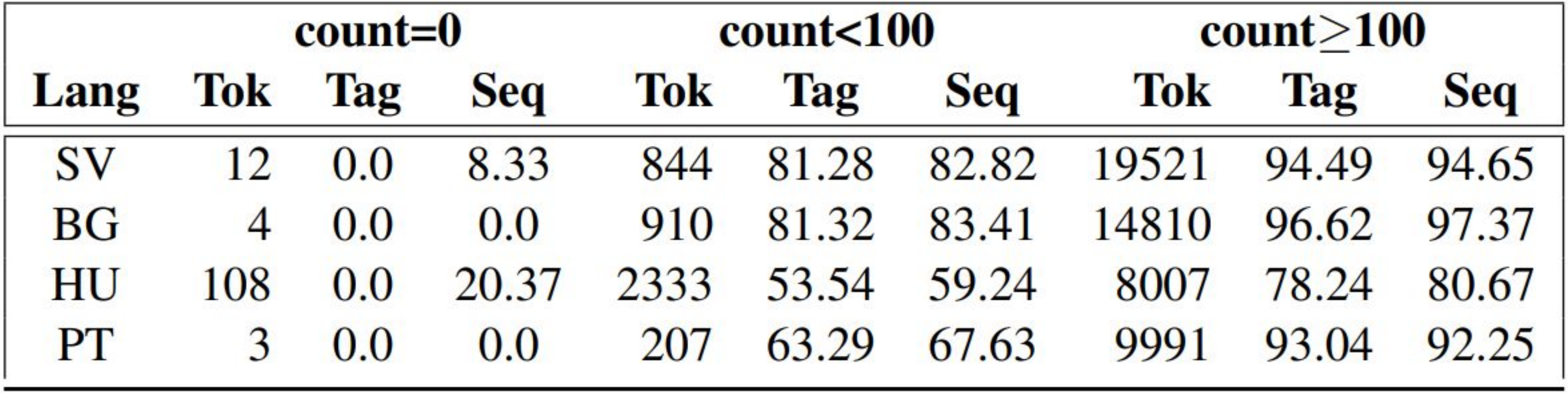


Method	A	U	T
word	94.38	98.70	96.72
word+context	96.21	98.52	97.69
word+context+output	96.43	98.80	97.79

A: ambiguous accuracy
U: unambiguous accuracy
T: total accuracy

Generating sequences vs whole tags

(sequences better, especially for rare tags)

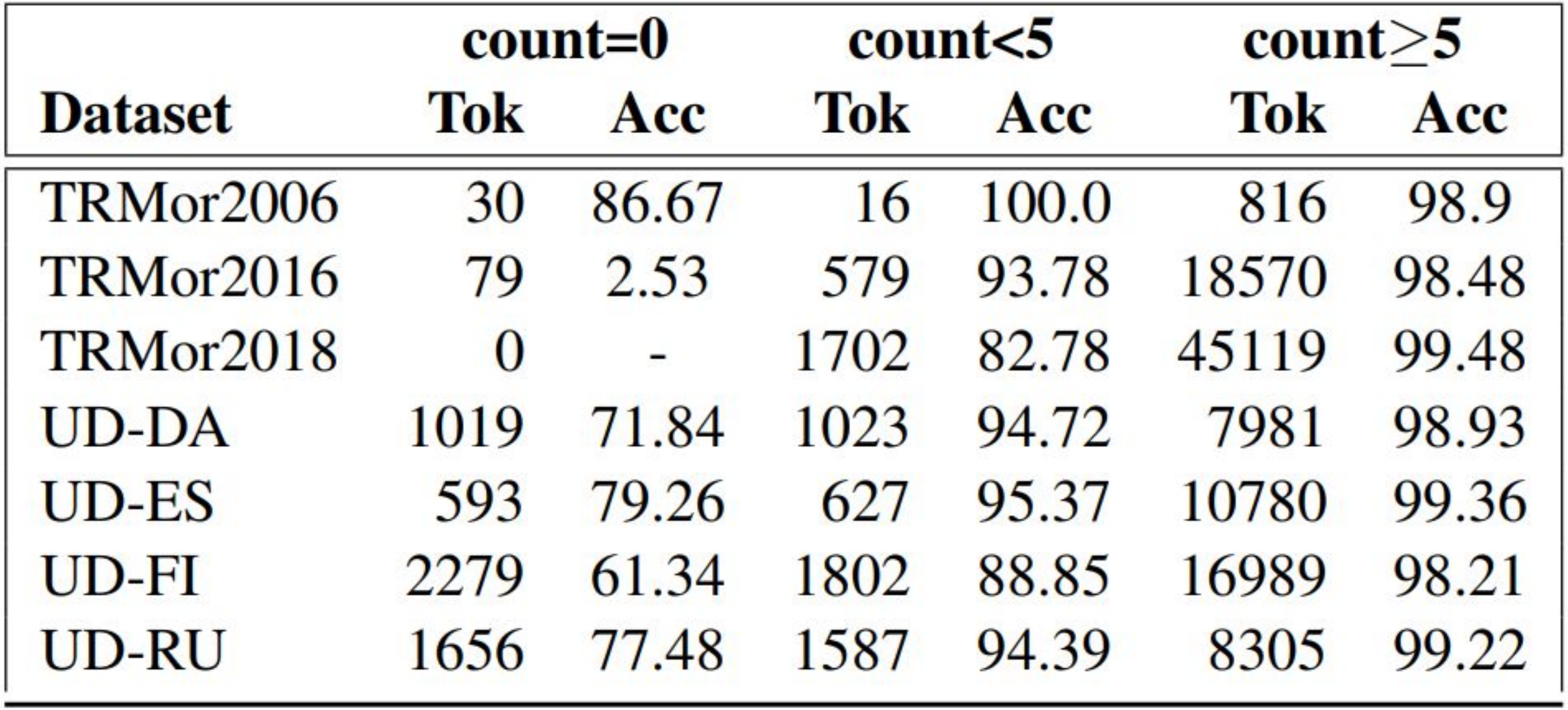


	count=0			count<100			count≥100		
Lang	Tok	Tag	Seq	Tok	Tag	Seq	Tok	Tag	Seq
SV	12	0.0	8.33	844	81.28	82.82	19521	94.49	94.65
BG	4	0.0	0.0	910	81.32	83.41	14810	96.62	97.37
HU	108	0.0	20.37	2333	53.54	59.24	8007	78.24	80.67
PT	3	0.0	0.0	207	63.29	67.63	9991	93.04	92.25

Tag: whole-tag generator
Seq: sequence decoder (Morse)

Generating rare lemmas

(Morse can generate lemmas it has never seen)



	count=0		count<5		count≥5	
Dataset	**Tok**	**Acc**	**Tok**	**Acc**	**Tok**	**Acc**
TRMor2006	30	86.67	16	100.0	816	98.9
TRMor2016	79	2.53	579	93.78	18570	98.48
TRMor2018	0	-	1702	82.78	45119	99.48
UD-DA	1019	71.84	1023	94.72	7981	98.93
UD-ES	593	79.26	627	95.37	10780	99.36
UD-FI	2279	61.34	1802	88.85	16989	98.21
UD-RU	1656	77.48	1587	94.39	8305	99.22

Code and data available

Code: https://github.com/ai-ku/Morse.jl

Data: https://github.com/ai-ku/TrMor2018

Multilingual data (UD v2.1):
https://universaldependencies.org