Part-of-Speech Tagging for Northern Kurdish

Peshmerge Morad p.morad@hotmail.com

Sina Ahmadi sina.ahmadi@uzh.ch

Lorenzo Gatti l.gatti@utwente.nl

UNIVERSITY OF TWENTE.

University of Zurich^UZH

Task

Build a POS Tagger for Northern Kurdish

Contributions

1. Training 7 different POS taggers for Northern Kurdish.
2. Augmenting and enriching the UD Kurmanji treebank for training purposes.
3. Creating a novel manually annotated and tokenized gold-standard dataset consisting of 136 sentences (2, 937 tokens) for testing purposes.
4. Demonstrating the effect of tokenization and various linguistic features of Northern Kurdish on the task of POS tagging.

Future Work

1. Working on syntactic parsing for Northern Kurdish.
2. Employment of LLMs or POS models from other closely related languages like Persian or dialects like Central Kurdish.
3. Examining the impact of our POS tagging models and annotation schemes on other downstream tasks.

Repository

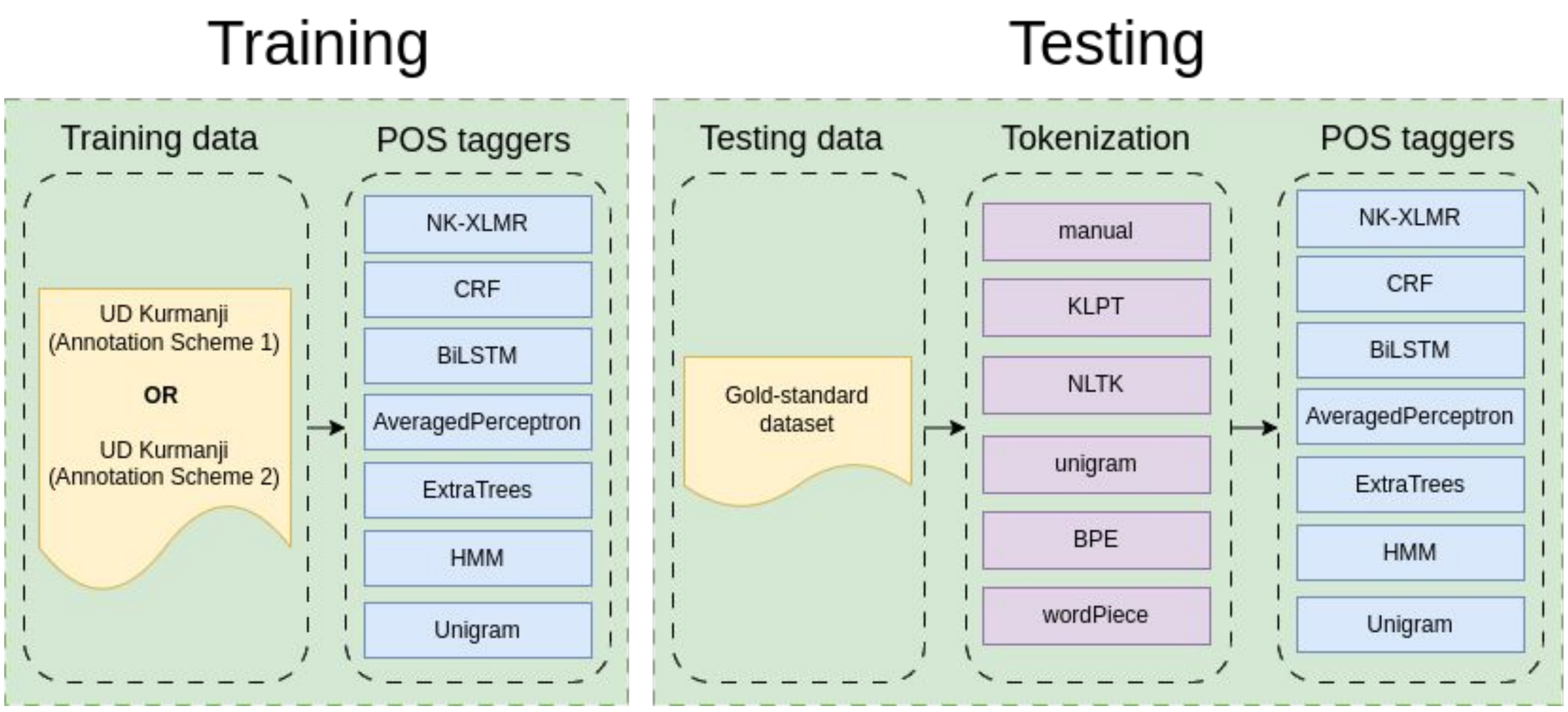
https://github.com/peshmerge/northern-kurdish-pos-tagging

Motivation

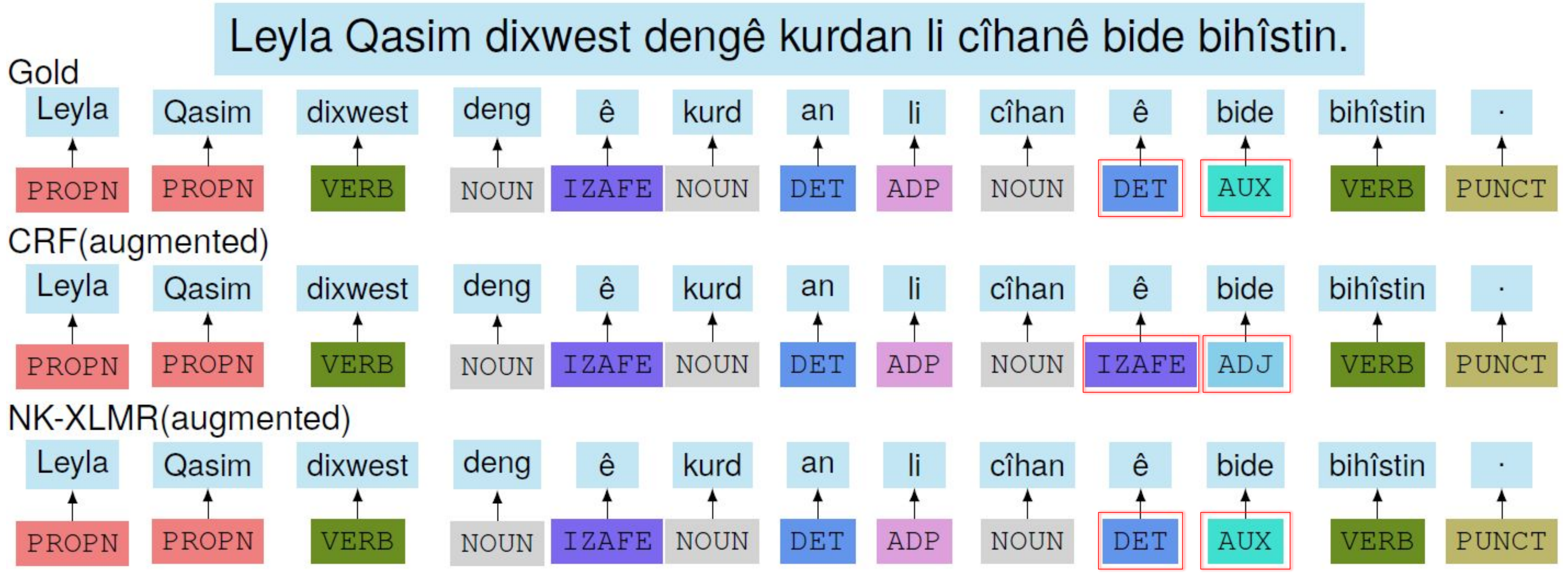
The Kurdish language belongs to the Northwestern Iranic branch within the Indo-European languages family, spoken by more than 30 million people.
Kurdish has 4 dialects and Northern Kurdish is the biggest one.

Except for the work of Walther et al., 2010, there has been no dedicated work for POS tagging for Northern Kurdish. In addition, available datasets like UD Kurmanji do not explicitly address the linguistic features such as the oblique and the construct (Izafe) markers.

Approach



Results



Leyla Qasim dixwest dengê kurdan li cîhanê bide bihîstin.

Gold

Leyla	Qasim	dixwest	deng	ê	kurd	an	li	cîhan	ê	bide	bihîstin	.
PROPN	PROPN	VERB	NOUN	IZAFE	NOUN	DET	ADP	NOUN	DET	AUX	VERB	PUNCT

CRF(augmented)

Leyla	Qasim	dixwest	deng	ê	kurd	an	li	cîhan	ê	bide	bihîstin	.
PROPN	PROPN	VERB	NOUN	IZAFE	NOUN	DET	ADP	NOUN	IZAFE	ADJ	VERB	PUNCT

NK-XLMR(augmented)

Leyla	Qasim	dixwest	deng	ê	kurd	an	li	cîhan	ê	bide	bihîstin	.
PROPN	PROPN	VERB	NOUN	IZAFE	NOUN	DET	ADP	NOUN	DET	AUX	VERB	PUNCT

Outputs of CRF, NK-XLMR (augmented) compared to the gold annotations for a sentence from the gold standard dataset.
Translation: 'Layla Qasim wanted to make the voice of the Kurds heard in the world.'

Model	F1	Acc
Baseline (Unigram)	0.4	0.51
HMM	0.37	0.46
ExtraTrees	0.41	0.52
AveragedPerceptron	0.44	0.54
BiLSTM	0.42	0.51
CRF	0.46	0.54
NK-XLMR	**0.57**	**0.62**

Results of our POS models trained on UD Kurmanji original and evaluated on the gold-standard dataset

Model	F1	Acc
Baseline (Unigram)	0.59	0.73
HMM	0.62	0.77
ExtraTrees	0.61	0.79
AveragedPerceptron	0.68	0.83
BiLSTM	0.72	0.83
CRF	0.74	0.84
NK-XLMR	**0.77**	**0.87**

Results of our POS models trained on UD Kurmanji augmented and evaluated on the gold-standard dataset