CMU-01 at SIGMORPHON: Morphological Analysis and Lemmatization in Context

Aditi Chaudhary, Elizabeth Salesky, Gayatri Bhat
David Mortensen, Jaime Carbonell, Yulia Tsvetkov
Carnegie Mellon University

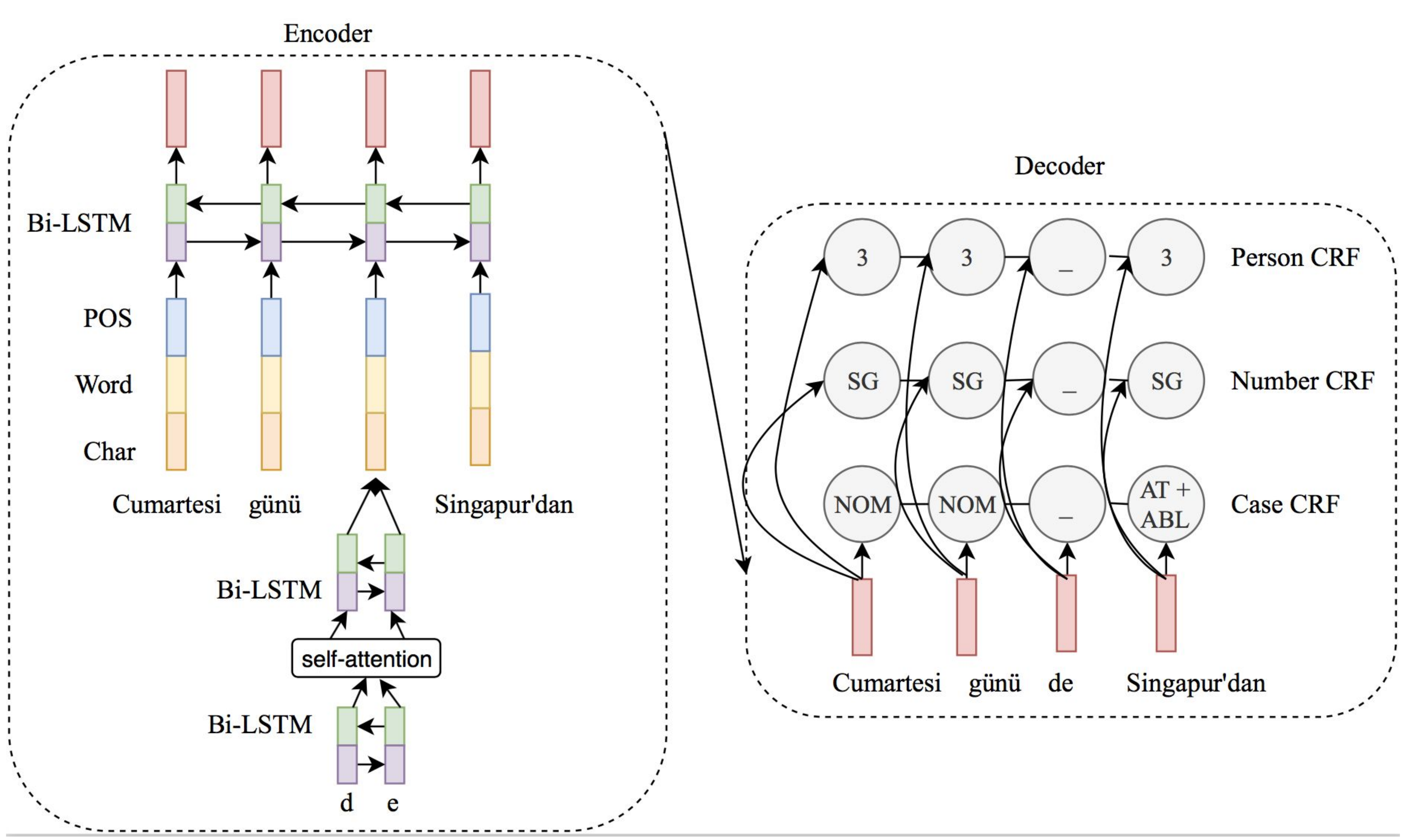
Carnegie Mellon University
School of Computer Science

Language Technologies Institute

Research Problem

- The SIGMORPHON 2019 Task 2 requires us to produce the lemma and morpho-syntactic description of each token in a sequence, for 107 treebanks.
- However, most treebanks are under-resourced, making it challenging to train deep neural models for them
- We approach this task with a hierarchical neural conditional random field (CRF) model which predicts each coarse-grained feature (eg. POS, Case, etc.) independently.
- To tackle the challenge of data-scarcity we propose a multi-lingual transfer training regime where we transfer from multiple related languages that share similar typology and/or orthography.

Model



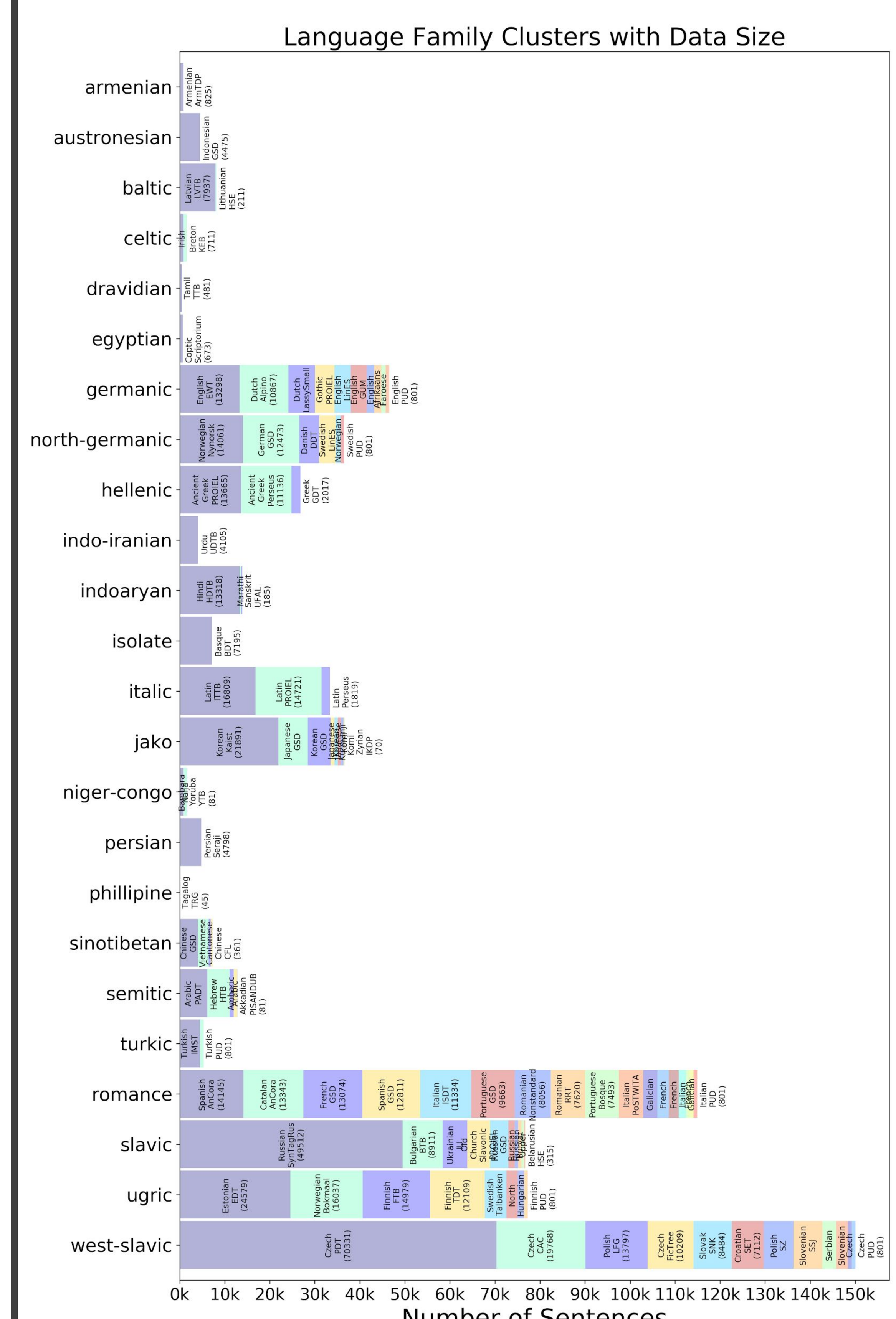
Hierarchical neural model for contextual morphological analysis with independent CRF decoders for each coarse-grained feature *F* such as Number, Person etc. POS embeddings are concatenated to the word and char-level representationse. This model has $|F|$*-1* decoders since POS tagger is run separately as a prior step.
MDCRF refers to the above model without POS embeddings having all $|F|$ decoders.

Training Data

- We model feature-wise prediction for each coarse-grained feature *F*={POS, Gender, …} and transform the dataset from UniMorph schema to key-value format. Therefore, we get:

N;PL;FEM ⇒ POS=N;Number=PL;Gender=FEM

- We augment training data using language clusters based on typographical and orthographic similarity, as shown below:



Results

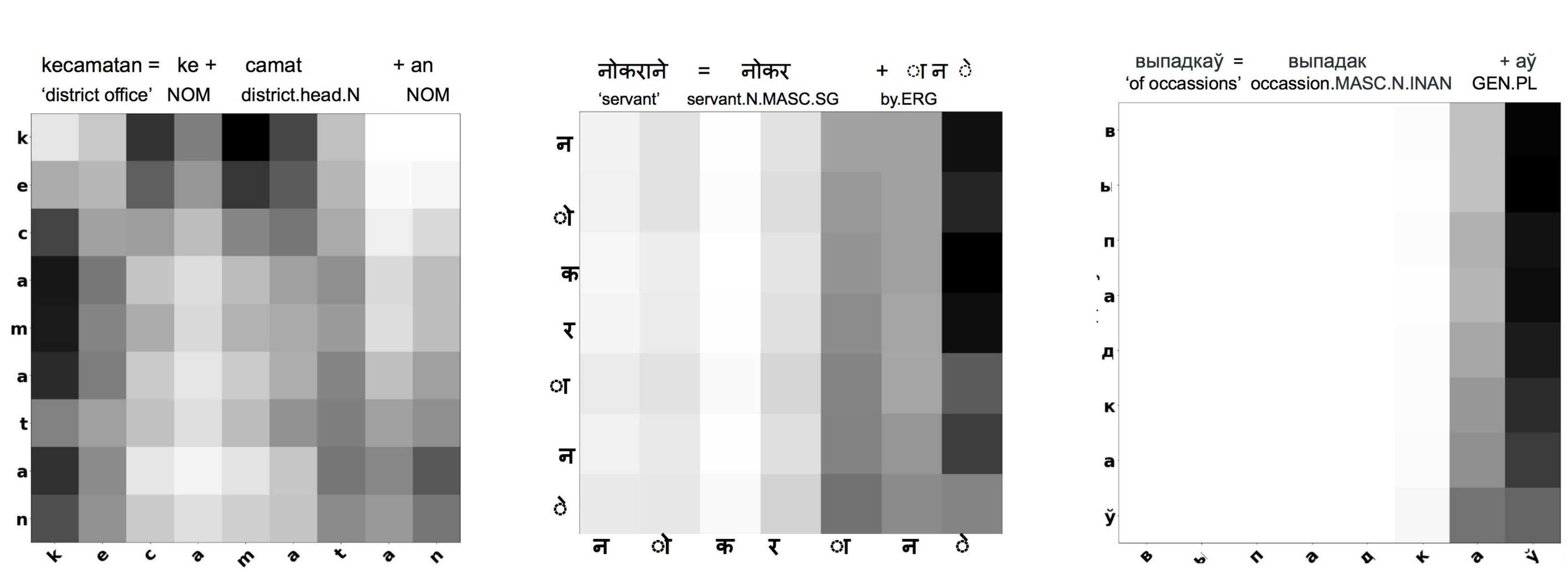
- Our system achieves an average improvement of **+15.30 (accuracy)** and **+4.93 (F1)** over the provided baseline. (McCarthy et al., 2019).
- We compare our model for bi-lingual transfer with previous bsselines:

Language	Model	tgt-size=100 Accuracy	tgt-size=100 F1-Macro	tgt-size=100 F1-Micro	tgt-size=1,000 Accuracy	tgt-size=1,000 F1-Macro	tgt-size=1,000 F1-Micro
RU/BG	MDCRF + POS + MULTI-SOURCE	**69.13**	**85.78**	**85.86**	**82.72**	**92.15**	**92.17**
	(Malaviya et al., 2018)	46.89	64.75	64.46	67.56	82.06	82.11
	(Cotterell and Heigold, 2017)	52.76	58.23	58.41	71.90	77.89	77.97
FI/HU	MDCRF + POS + MULTI-SOURCE	**57.32**	**80.11**	**78.86**	**70.24**	**85.44**	**84.86**
	(Malaviya et al., 2018)	45.41	68.63	68.07	63.93	85.06	84.12
	(Cotterell and Heigold, 2017)	51.74	68.15	66.82	61.8	75.96	76.16

Q. What is the model learning?

- Character-level attention maps for three typologically different languages:

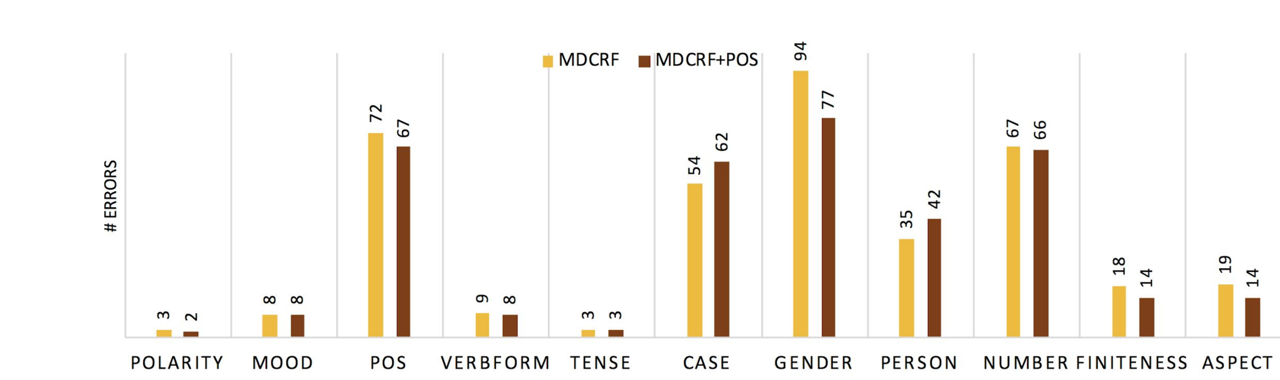
INDONESIAN MARATHI BELARUSIAN



- Marathi and Belarusian display morphological inflections as suffix.
- Indonesian displays inflections in the form of prefix, suffix and circumfix.

Q. Why does adding POS help?

- Removing POS gives a drop in accuracy which is significant for low-medium resource languages: **Maratih (-6.12 acc), Ukrainian (-3.57 acc).**
- Number of errors per feature *F*: POS helped reduced Gender errors in Marathi

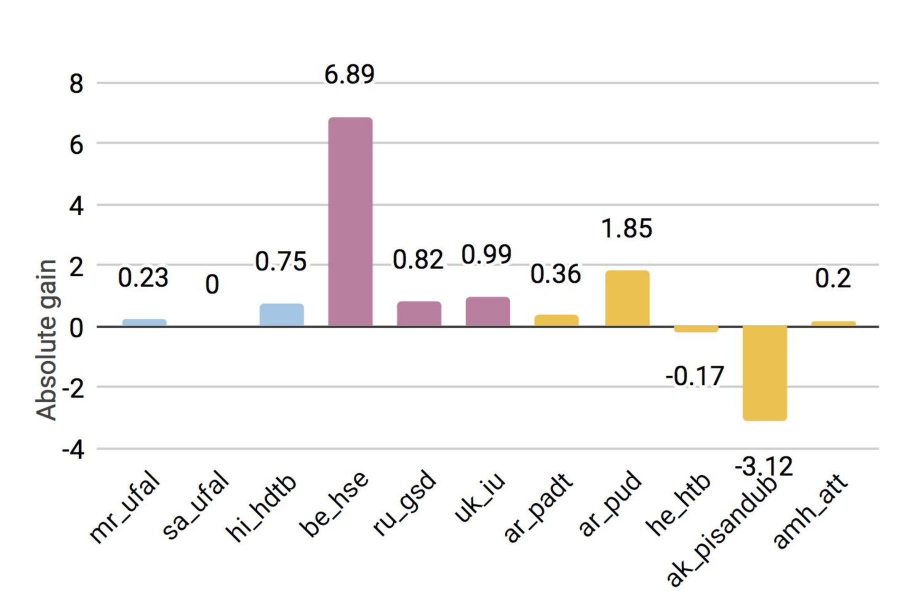


	POLARITY	MOOD	POS	VERBFORM	TENSE	CASE	GENDER	PERSON	NUMBER	FINITENESS	ASPECT
MDCRF	4	4	348	16	14	121	198	91	243	65	49
MDCRF+POS	3	6	348	17	16	121	197	106	248	78	63

- Eg. For some words gender may be inferred from inflectional form, but for others such as **(किंमत N.FEM.SG.ACC)** the traditional female suffix ई is not present, in which casen having the POS information helps.

Q. Does time-depth matter for transfer learning?

- *Time-depth* is period of time that has elapsed since all languages in the group were a single language.
- We study the following clusters: Indo-Aryan, Slavic and Semitic.



- Transfer helps most for Slavic (+2.9 accuracy), and next for Indo-Aryan cluster (+0.32 accuracy). There is a negative effect for the Semitic cluster (-0.0176 accuracy).
- Indo-Aryan and Slavic have shallower time-depths: < 1000 years ⇒ better transfer!

Code: https://github.com/Aditi138/MorphologicalAnalysis
Contact: aschaudh@andrew.cmu.edu