Google AI

Morphosyntactic Tagging with a Meta-BiLSTM Model over Context Sensitive Token Encodings

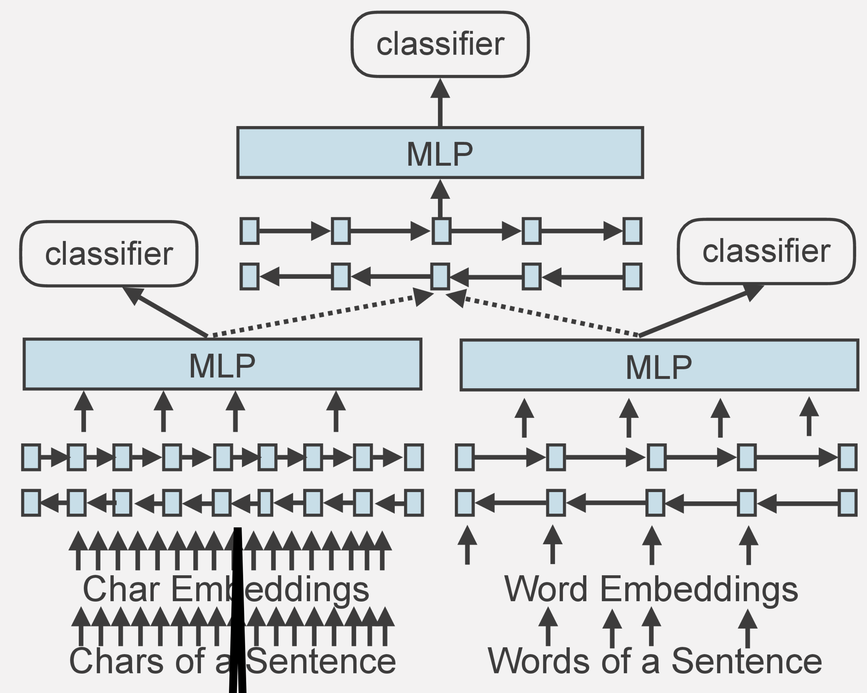
Bernd Bohnet, Ryan McDonald, Gonçalo Simões, Daniel Andor, Emily Pitler, Joshua Maynez

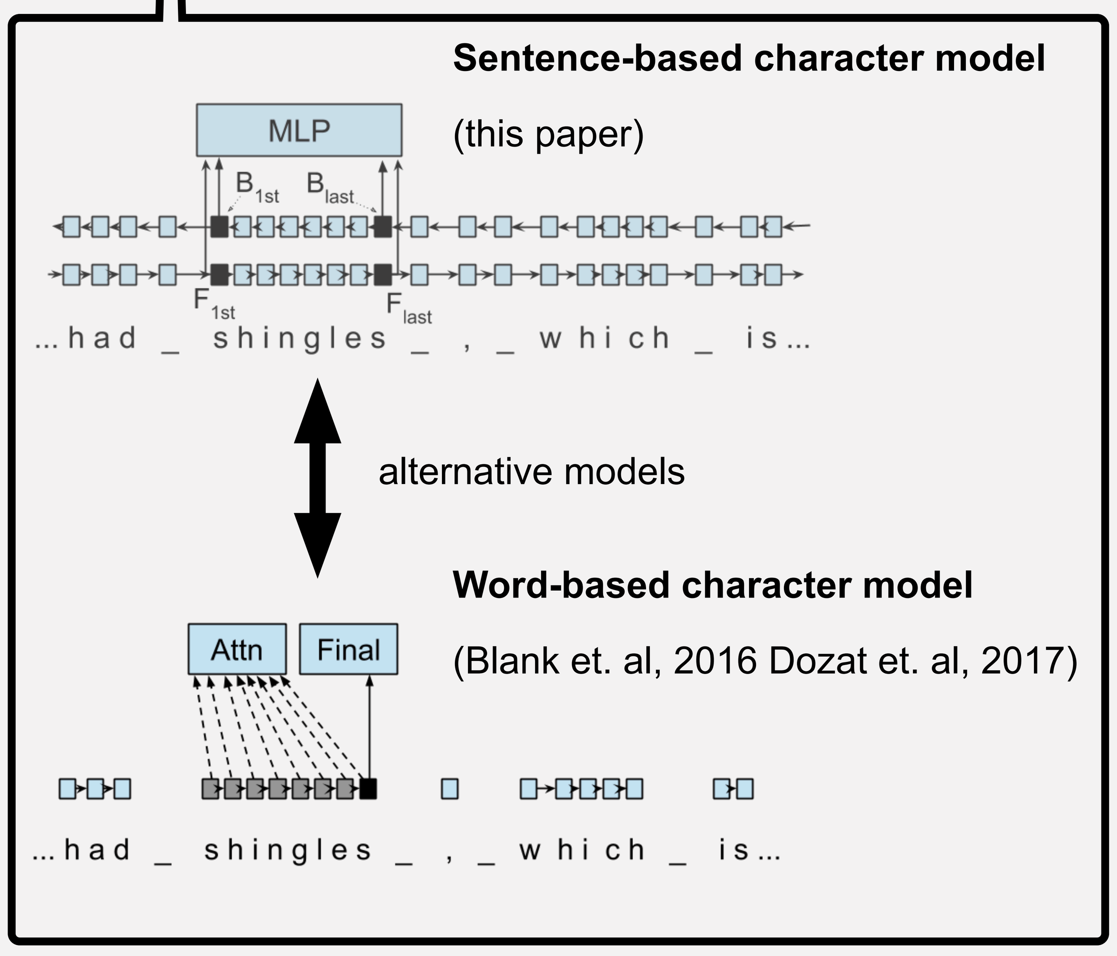
The problem, we address - sequence tagging:

```
I     had shingles , which is  a  painful disease .
PRP   VBD   NN*    ,  WDT  VBZ DET  JJ       NN    .
```

*Context insensitive character representations may have learned that unknown words ending with 's' might be plural nouns (NNS).

System architecture





Ablation Studies

Character and Word Model Contribution

F1 score for the character, word, meta models, and standard deviations of 10 random restarts.

dev. set lang.	num. exp.	mean char	mean word	mean meta	stdev char	stdev word	stdev meta
bg	10	96.59	93.20	**97.04**	0.05	0.11	0.04
el	10	96.43	95.36	**97.01**	0.13	0.11	0.09
grc	10	88.28	73.52	**88.85**	0.21	0.29	0.22
la_ittb	10	91.45	87.98	**91.94**	0.14	0.30	0.05
en	10	95.10	94.77	**95.88**	0.14	0.08	0.05
ru	10	95.98	93.50	**96.61**	0.06	0.17	0.07
tr	10	93.77	90.48	**94.67**	0.11	0.33	0.14

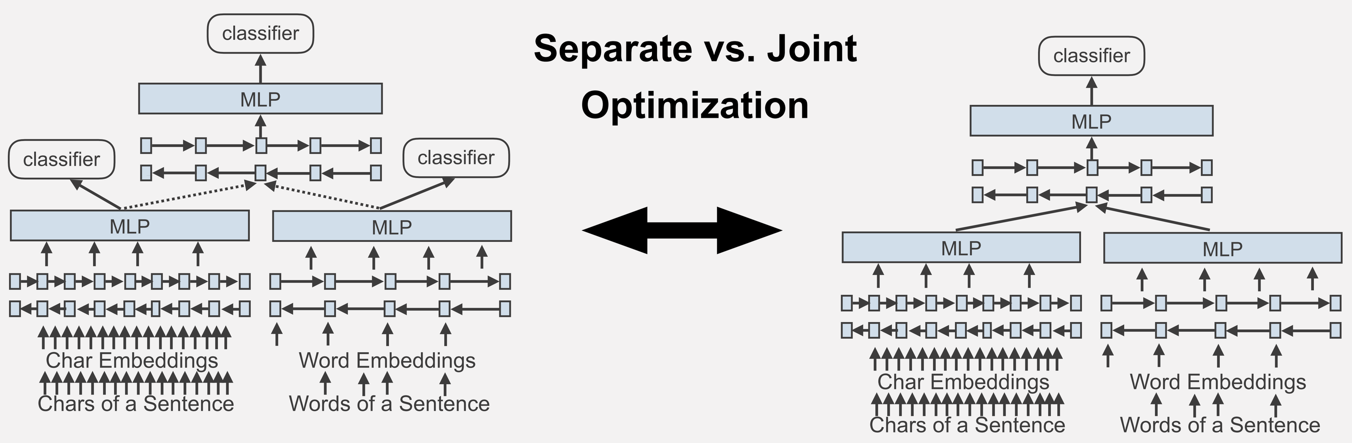
XPOS F1 Scores.

Word Character vs Sentence Char Model

F1 score for the word-based versus sentence-based character model and optimized separately.

dev. set	word char model	sentence char model
el	89.05	93.41
la_ittb	93.22	95.69
ru	88.94	92.31
tr	87.78	90.77

➤ Sentence-based character model is more accurate than word-based character model especially when optimized separately.



Separate vs. Joint Optimization

We compare jointly training of the character model, word model, and Meta-BiLSTM versus training each model separately.

Optimization	Avg. F1 Score morphology	Avg. F1 Score xpos
separate	**94.57**	**94.85**
jointly	94.15	94.48

➤ Separate optimization of the word, character and meta model is more accurate on average than full back-propagation using a single loss function.

Part-of-Speech

lang.	CONLL Winner	DQM	ours	RRIE
cs cac	95.16	95.16	**96.91**	36.2
cs	95.86	95.86	**97.28**	35.5
fi	97.37	97.37	**97.81**	16.7
sl	94.74	94.74	**95.54**	15.2
la ittb	94.79	94.79	**95.56**	14.8
grc	84.47	84.47	**86.51**	13.1
bg	96.71	96.71	**97.05**	10.3
ca	98.58	98.58	**98.72**	9.9
grc proiel	97.51	97.51	**97.72**	8.4
pt	83.04	83.04	**84.39**	8.0
cu	96.20	96.20	**96.49**	7.6
it	97.93	97.93	**98.08**	7.2
fa	97.12	97.12	**97.32**	6.9
ru	96.73	96.73	**96.95**	6.7
sv	96.40	96.40	**96.64**	6.7
ko	93.02	93.02	**93.45**	6.2
sk	85.00	85.00	**85.88**	5.9
nl	90.61	90.61	**91.10**	5.4
fi ftb	95.31	95.31	**95.56**	5.3
de	97.29	97.29	**97.39**	4.7
tr	93.11	93.11	**93.43**	4.6
hi	97.01	97.01	**97.13**	4.0
es ancora	98.73	98.73	**98.78**	3.9
ro	96.98	96.98	**97.08**	3.6
la proiel	96.93	96.93	**97.00**	2.3
pl	91.97	91.97	**92.12**	1.9
ar	87.66	87.66	**87.82**	1.3
gl	97.50	97.50	**97.53**	1.2
sv lines	94.84	94.84	**94.90**	1.2
cs clt	89.98	89.98	**90.09**	1.1
lv	80.05	80.05	**80.20**	0.8
zh	**88.40**	85.07	85.10	0.2
en lines	**95.41**	**95.41**	95.39	-0.4
ur	**92.30**	**92.30**	92.21	-1.2
he	**83.24**	82.45	82.16	-1.7
vi	**75.42**	73.56	73.12	-1.7
gl treegal	**91.65**	**91.65**	91.40	-3.0
en	**94.82**	**94.82**	94.66	-3.1
en partut	**95.08**	**95.08**	94.81	-5.5
pt br	**98.22**	**98.22**	98.11	-6.2
et	**95.05**	**95.05**	94.72	-6.7
el	**97.76**	**97.76**	97.53	-10.3
macro-avg	93.18	93.11	**93.40**	-

CoNLL 17 Winner and DQM: Dozat et al., 2017 and our system.

System	Accuracy
Søgaard (2011)	97.50
Huang et al. (2015)	97.55
Choi (2016)	97.64
Andor et al. (2016)	97.44
Dozat et al. (2017)	97.41
ours	**97.96**

Accuracy on Wall Street

Morphological Features

lang.	CONLL Winner	DQM Reimpl.	ours	RRIE
cs cac	90.72	94.66	**96.41**	27.9
ru syn.	94.55	96.70	**97.53**	23.1
cs	93.14	96.32	**97.14**	22.3
la ittb	94.28	96.45	**97.12**	18.9
sl	90.08	95.26	**96.03**	16.2
ca	97.23	97.85	**98.13**	13.0
fi ftb	93.43	95.96	**96.42**	11.4
no bok.	95.56	96.95	**97.26**	10.2
grc proiel	90.24	91.35	**92.22**	10.1
fr sequoia	96.10	96.62	**97.62**	10.1
la proiel	89.22	91.52	**92.35**	9.8
es ancora	97.72	98.15	**98.32**	9.7
da	94.83	96.62	**96.94**	9.5
fi	92.43	94.29	**94.83**	9.5
sv	95.15	96.52	**96.84**	9.2
pt	94.62	95.89	**96.27**	9.2
grc	88.00	90.39	**91.13**	9.0
no nyn.	95.25	96.79	**97.08**	9.0
de	83.11	89.78	**90.70**	9.0
ru	87.27	91.99	**92.69**	8.7
hi	91.03	90.72	**91.78**	8.1
cu	88.90	88.93	**89.82**	8.0
fa	96.34	97.23	**97.45**	7.9
tr	87.03	89.39	**90.21**	7.7
en partut	92.69	93.93	**94.40**	7.7
sk	81.23	87.54	**88.48**	7.5
eu	89.57	92.48	**93.04**	7.4
pt br	99.73	99.73	**99.75**	7.4
es	96.34	96.42	**96.68**	7.3
ko	99.41	99.44	**99.48**	7.1
ar	87.15	85.45	**88.29**	6.7
it	97.37	97.72	**97.86**	6.1
nl lassy	97.55	98.04	**98.15**	5.2
nl	90.04	92.06	**92.47**	5.2
pl	86.53	91.71	**92.14**	5.2
ur	81.03	83.16	**84.02**	5.1
bg	96.47	97.71	**97.82**	4.8
hr	85.82	90.64	**91.50**	3.8
he	**85.06**	79.34	79.76	2.0
et	84.62	88.18	**88.25**	0.6
zh	**92.90**	87.67	87.74	0.6
vi	**86.92**	82.23	82.30	0.4
ja	**96.84**	89.65	89.66	0.1
cs cltt	87.88	**90.41**	90.36	-0.5
lv	84.14	**87.00**	86.92	-0.6
el	91.37	**94.00**	93.92	-1.3
hu	72.61	**82.67**	82.44	-1.3
en	94.49	**95.93**	95.71	-5.4
macro-avg	91.51	92.89	**93.31**	-

CoNLL 17 Winner and DQM Reimpl.: Dozat et al., 2017 and our system.

Conclusions

- State-of-the-Art accuracy for a large number of treebanks for XPOS tags, morphological features and WSJ XPOS.
- Sentence-based character context provides substantial improvements especially for morphological rich languages.
- Meta tagger architecture allows to optimize the character and word model individually.

CoNLL 2018

- Used by the Uppsala team which **ranked 1st** for UPOS and morphology.

Download (open source):

https://github.com/google/meta_tagger

Hyperparameter Selection

With larger network sizes the capacity of the network increases, on the other hand it is prone to overfitting.

➤ For the BiLSTMs, we use 3 layers with 500 cells but for some languages much smaller settings yield to competitive results, e.g. for Vietnamese.

Grid search: Vietnamese