Heidelberg-Boston @ SIGTYP 2024 Shared Task: Enhancing Low-Resource Language Analysis With Character-Aware Hierarchical Transformers

Frederick Riemenschneider
riemenschneider@cl.uni-heidelberg.de

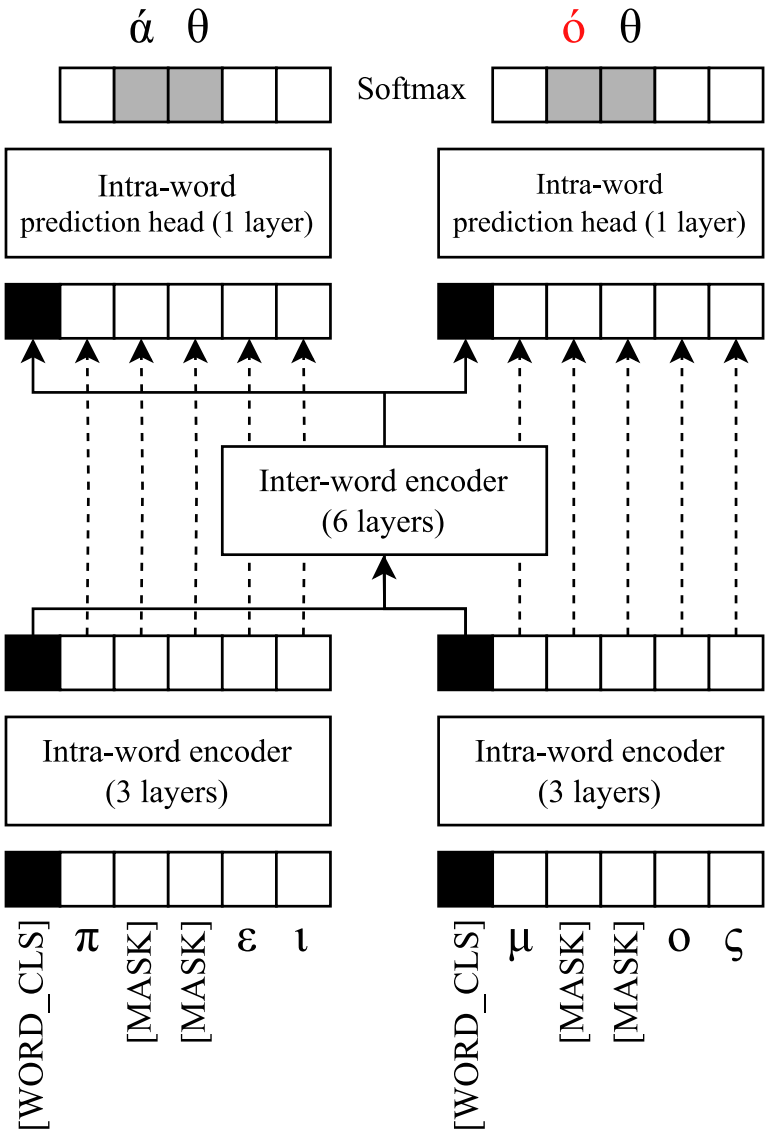
Kevin Krahn
kevin.krahn24@sattler.edu

March 2024

Tokenization

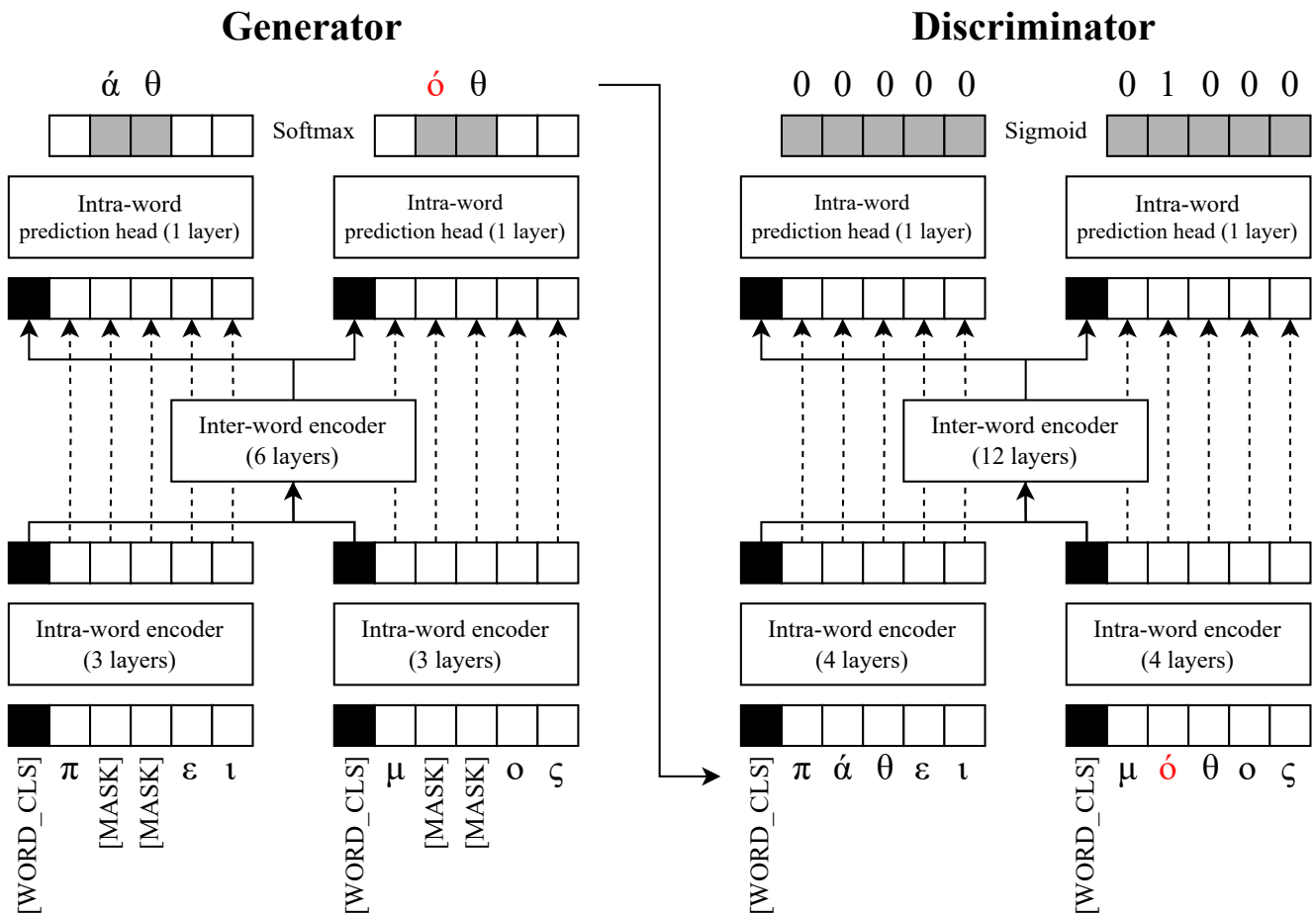
- **word-level**
  - ✗ low-resource context → rare occurrence of words
  - ✗ inflected languages → unique word forms exceedingly rare
- **subword-level**
  - ✗ much character information lost
  - ✗ no explicit word representations
- **character-level**
  - ✗ long sequence lengths
  - ✗ no explicit word representations

Hierarchical Language Model[1]



[1] Sun et al. 2023.

Replaced Token Detection



Masking Strategy

- **whole-word** masking
- **character** masking
- **character n-gram** masking

Lemmatization

- **sequence-to-sequence** task
- **character-based T5** model

Morphological Tagging

- **concatenate** intra- and inter-word embeddings
- **classifier** for each feature

$$\mathcal{L}_{\text{morph}} = \frac{1}{k} \sum_{m=0}^{k-1} \mathcal{L}_m$$

PoS Tagging

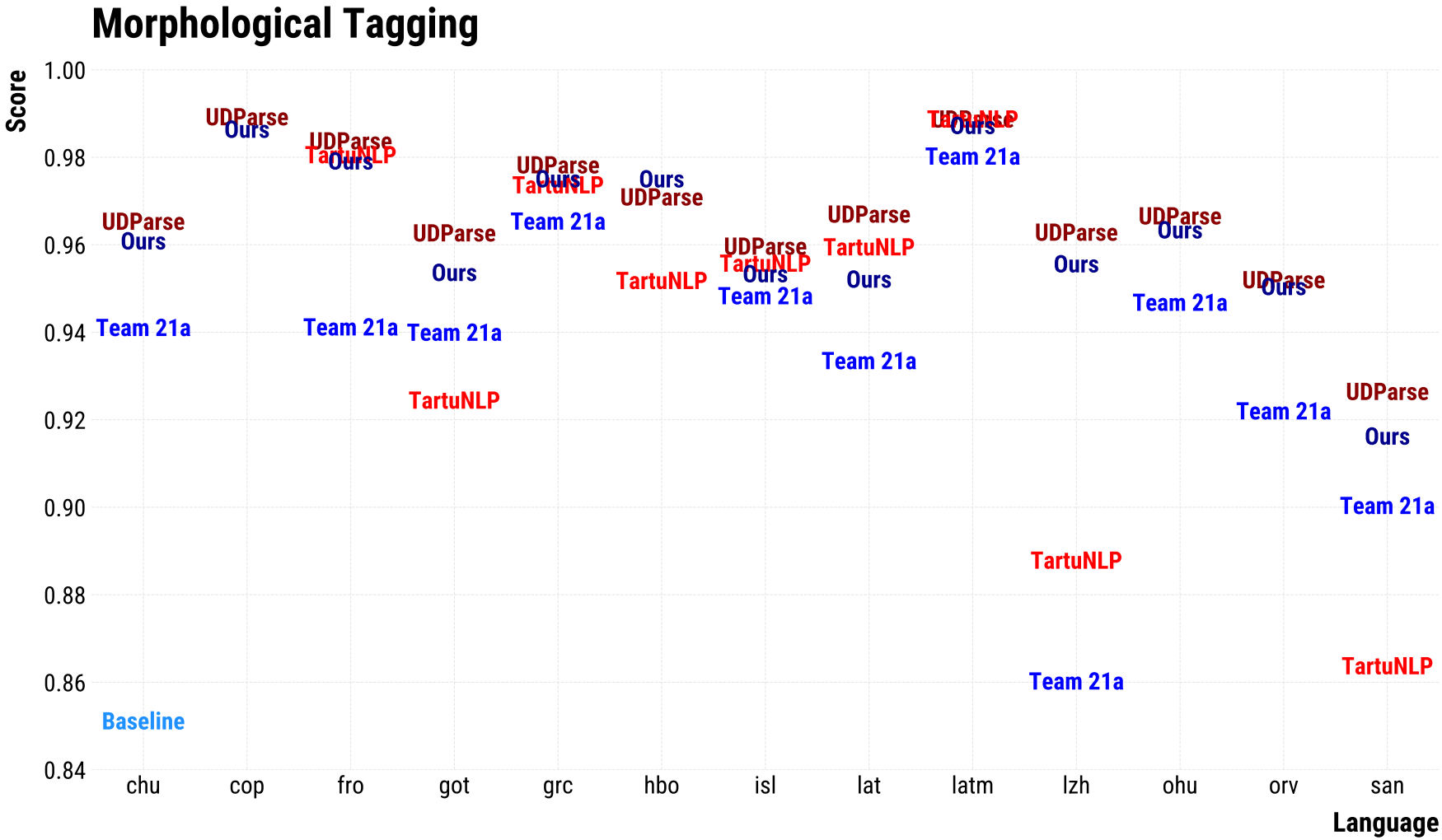
- **concatenate** intra- and inter-word embeddings
- **multi-task** learning with POS + morph

$$\mathcal{L}_{\text{UPoS}} + \mathcal{L}_{\text{morph}}$$

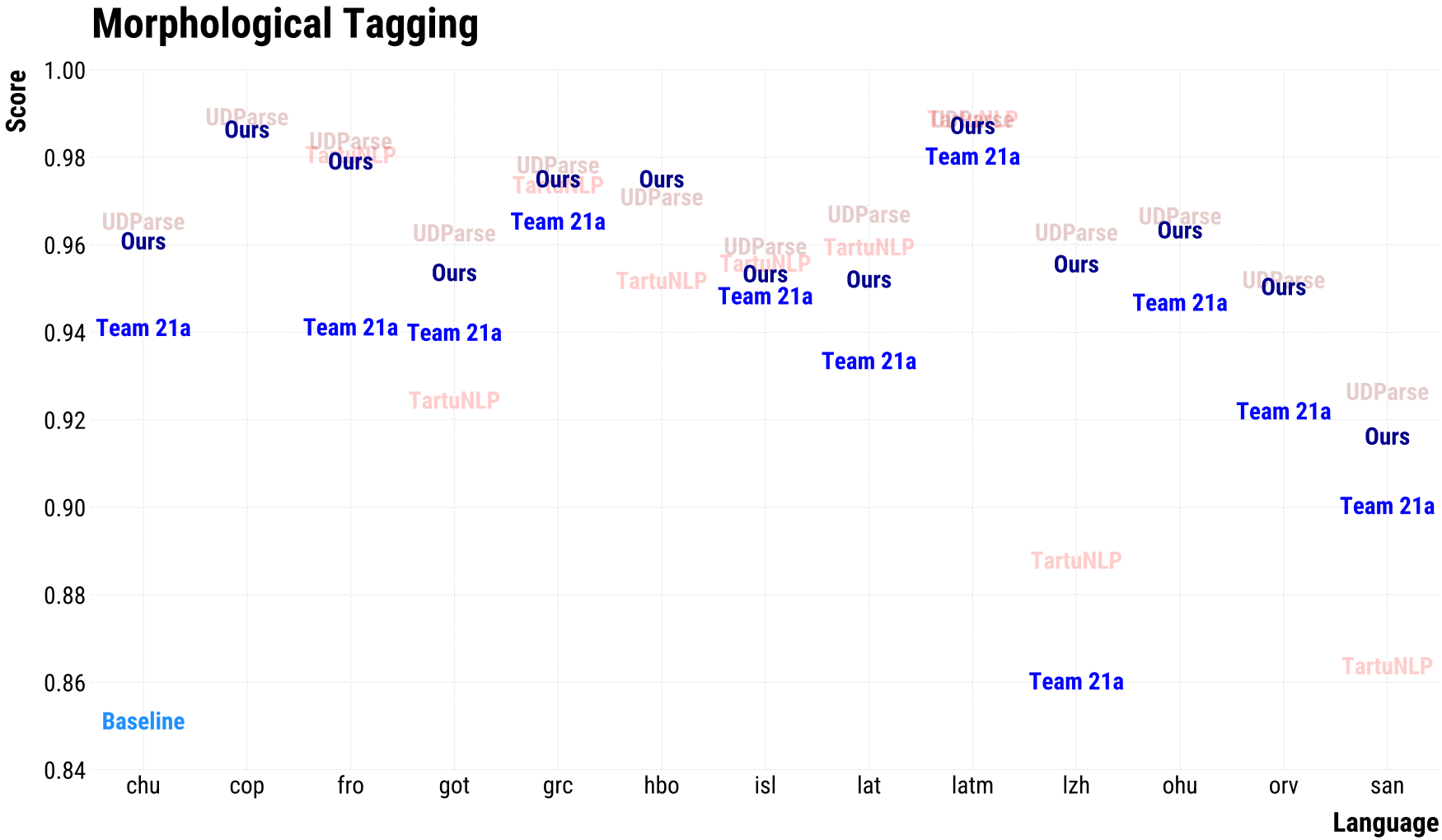
Lemmatization

Input	Label
quem PRON	quis
me PRON	ego
arbitramini VERB	arbitror
esse AUX	sum
non ADV	non
sum AUX	sum
ego PRON	ego

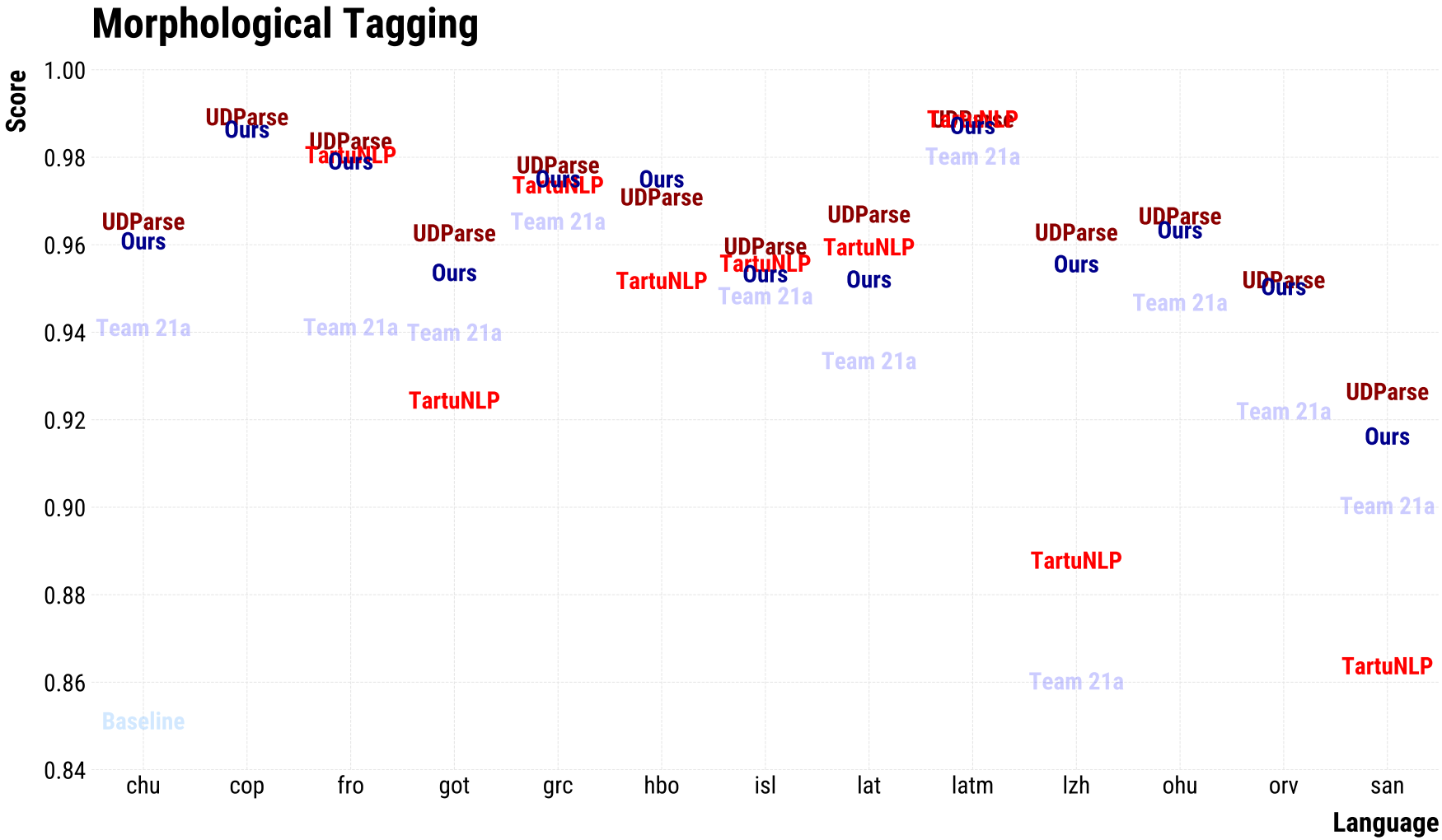
Results


Morphological Tagging
Score
1.00
0.98
0.96
0.94
0.92
0.90
0.88
0.86
0.84
chu
cop
fro
got
grc
hbo
isl
lat
latm
lzh
ohu
orv
san
Language
UDParse
Ours
Team 21a
TartuNLP
Baseline

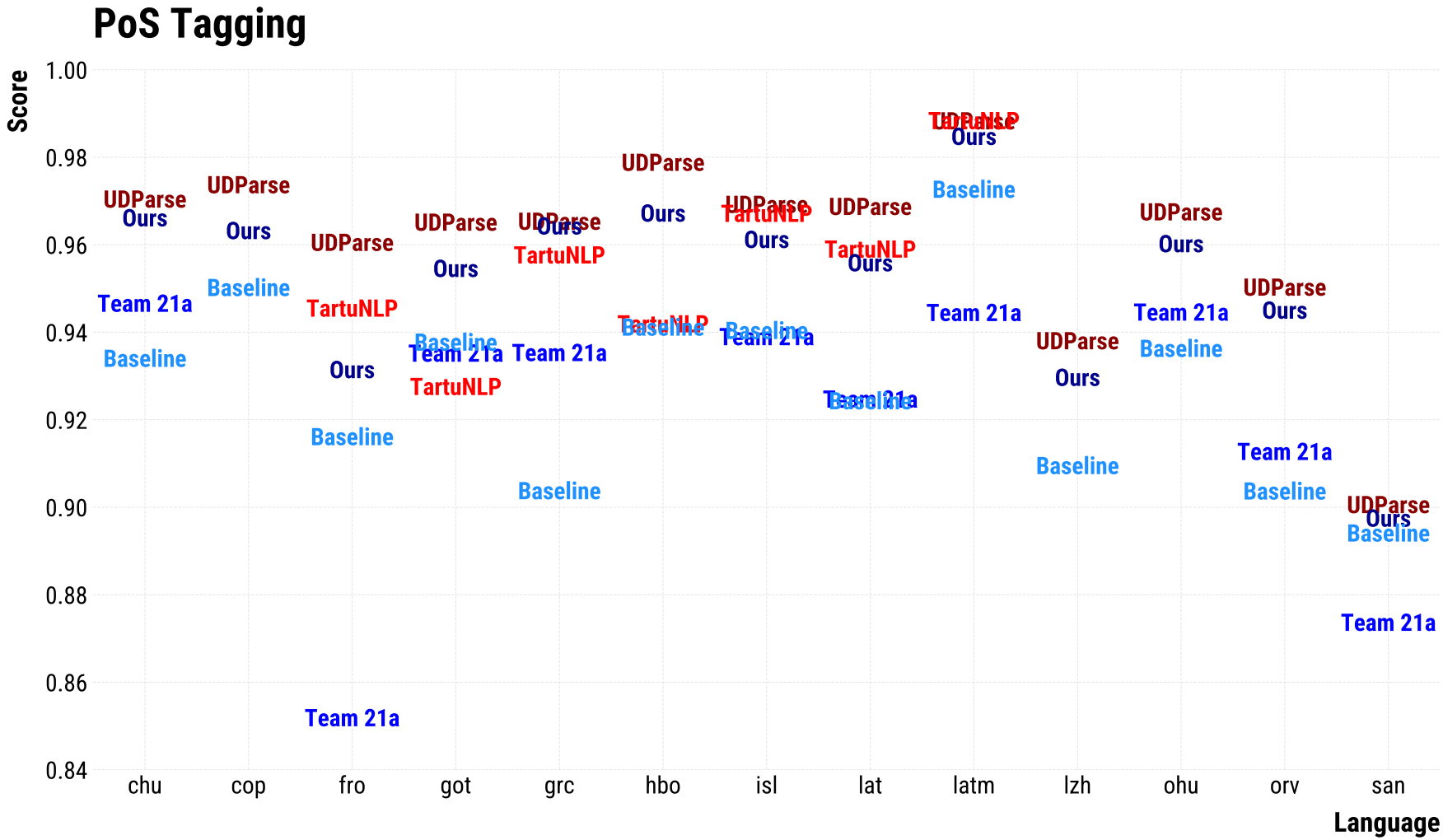
Results


Morphological Tagging
Score
1.00
0.98
0.96
0.94
0.92
0.90
0.88
0.86
0.84
chu
cop
fro
got
grc
hbo
isl
lat
latm
lzh
ohu
orv
san
Language
Baseline
Team 21a
Ours
TartuNLP
UDParse

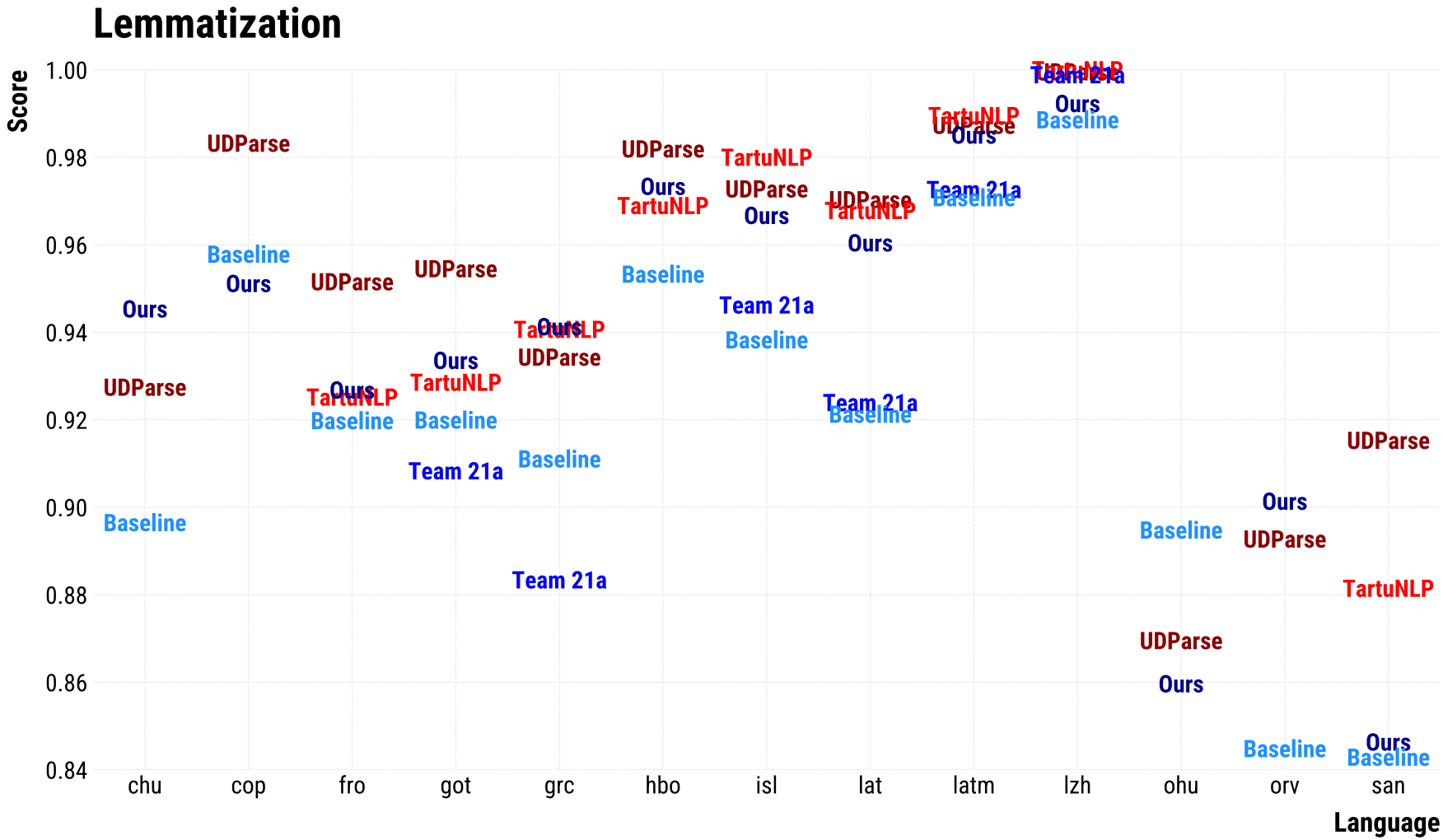
Results


Morphological Tagging
Score
1.00
0.98
0.96
0.94
0.92
0.90
0.88
0.86
0.84
chu
cop
fro
got
grc
hbo
isl
lat
latm
lzh
ohu
orv
san
Language
UDParse
Ours
TartuNLP
Team 21a
Baseline

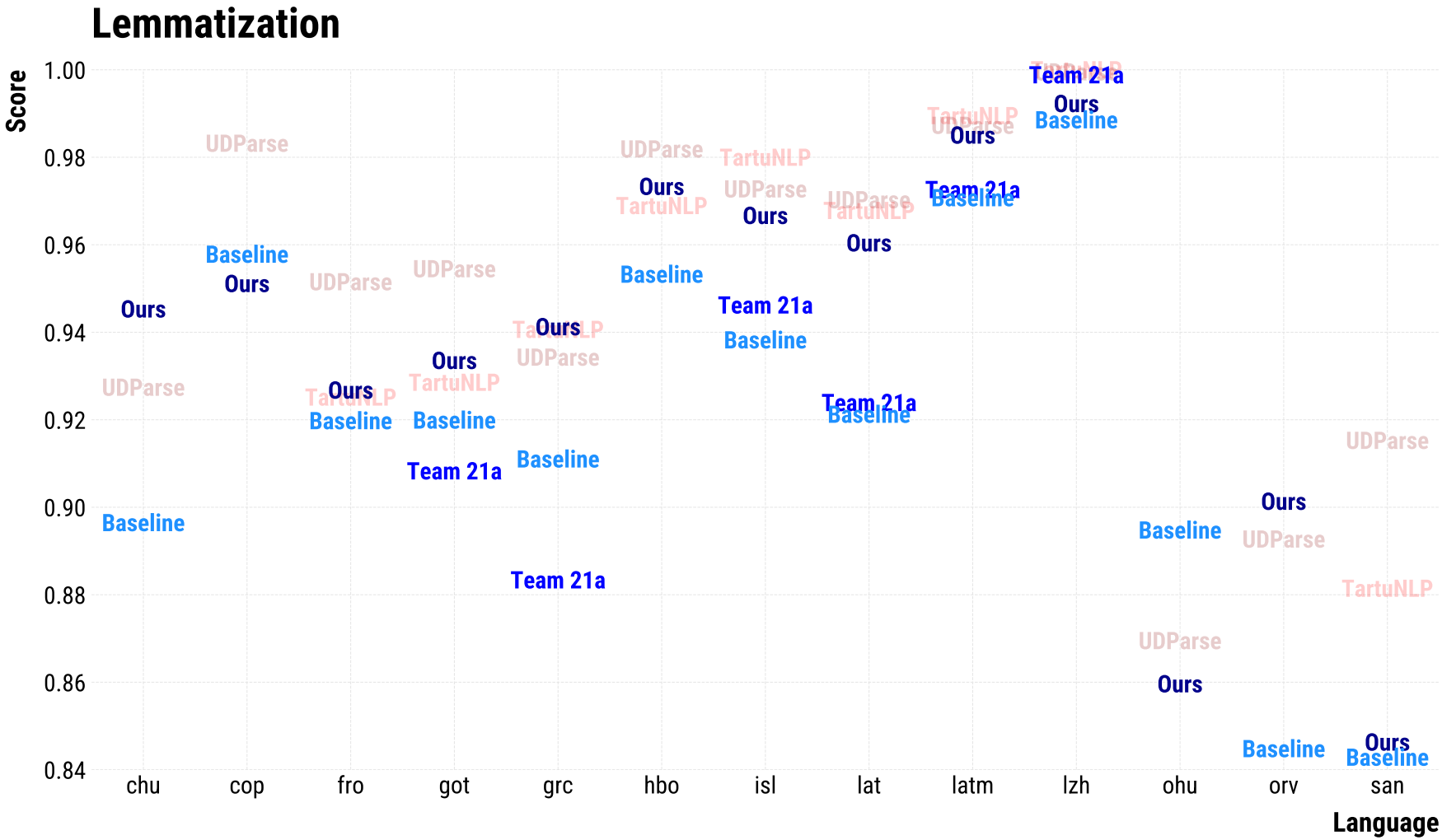
Results


PoS Tagging
Score
1.00
0.98
0.96
0.94
0.92
0.90
0.88
0.86
0.84
chu
cop
fro
got
grc
hbo
isl
lat
latm
lzh
ohu
orv
san
Language
UDParse
Ours
Baseline
Team 21a
TartuNLP

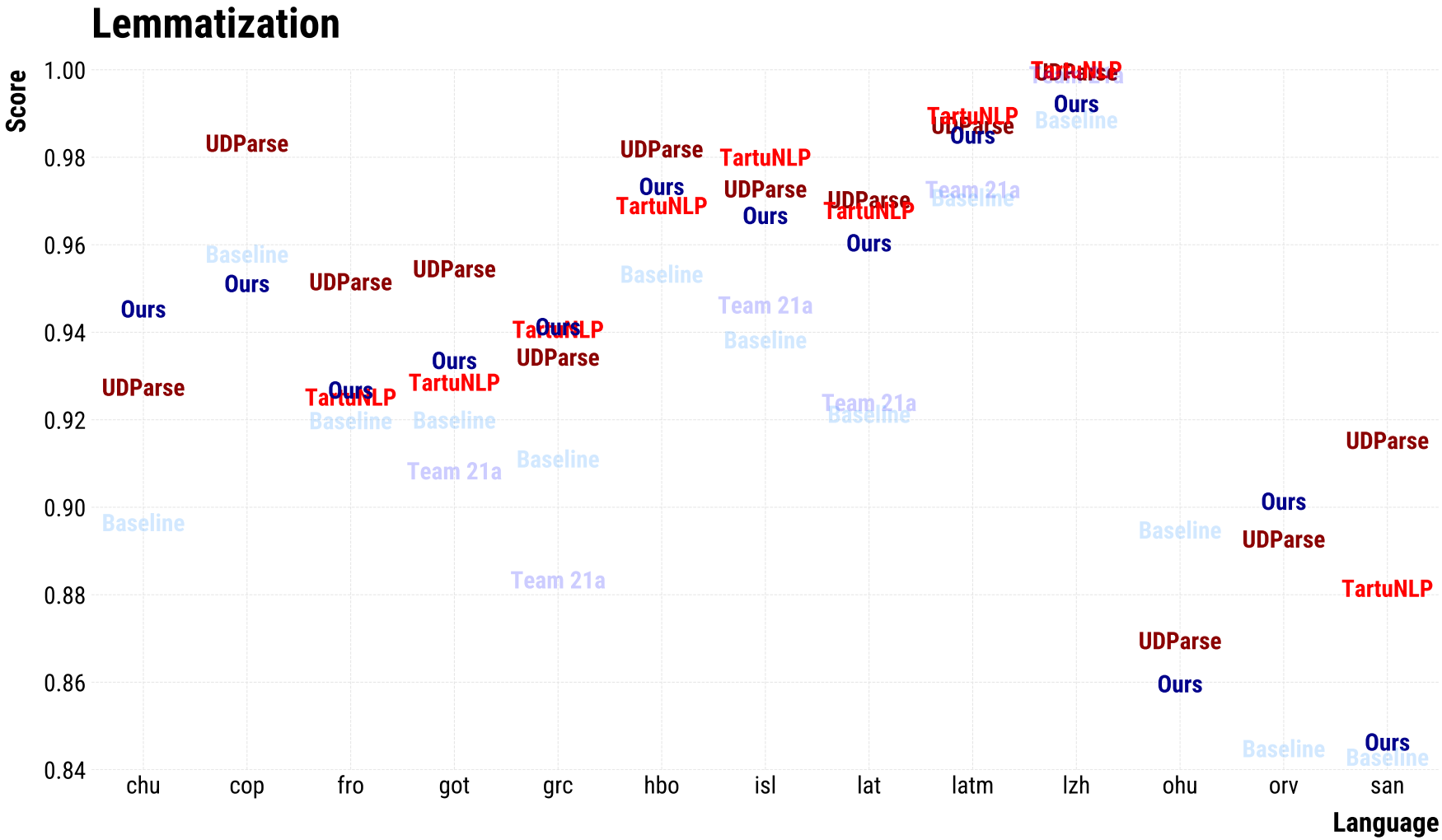
Results


Lemmatization
Score
1.00
0.98
0.96
0.94
0.92
0.90
0.88
0.86
0.84
chu
cop
fro
got
grc
hbo
isl
lat
latm
lzh
ohu
orv
san
Language
UDParse
Baseline
Ours
TartuNLP
Team 21a

Results


Lemmatization
Score
1.00
0.98
0.96
0.94
0.92
0.90
0.88
0.86
0.84
chu
cop
fro
got
grc
hbo
isl
lat
latm
lzh
ohu
orv
san
Language
Team 21a
Ours
Baseline
UDParse
TartuNLP

Results


Lemmatization
Score
1.00
0.98
0.96
0.94
0.92
0.90
0.88
0.86
0.84
chu
cop
fro
got
grc
hbo
isl
lat
latm
lzh
ohu
orv
san
Language
UDParse
Ours
TartuNLP
Baseline
Team 21a

Negative Results

- **multi-task** learning
- **tall** models
  - narrower and deeper architecture
  - benefits for Masked Langugage Modeling but not for Replaced Token Detection

Take-aways

- **Hierarchical Language Model** to learn **efficiently** from **every character**
- **DeBERTa-V3** with **Replaced Token Detection**
- **lemmatization** as **character-level sequence-to-sequence** problem

Thank you for your attention!