Toward Robust and Human-Like IGT Generation Models for Real-World Usage

Michael Ginn & Alexis Palmer

University of Colorado **Boulder**

Background

- **Palmer & Baldridge (2009)** and **Palmer et al. (2010)** finds that automated generation of interlinear glosses can significantly reduce annotator effort
- Recent approaches (**Moeller and Hulden, 2018**; **McMillan-Major, 2020**; **Zhao et al., 2010**) use various **statistical** and **neural methods** to automatically predict glosses with high accuracy

2023 SIGMORPHON Shared Task

Ginn et al., 2023

2023 SIGMORPHON Shared Task

- First public task for IGT glossing models
- Participants built systems for predicting glosses given transcriptions and (in some cases) translations

- **Open track:** gold standard segmentations, additional resources allowed
- **Closed track:** direct glossing from unsegmented words

2023 SIGMORPHON Shared Task

Languages

Arapaho
175k tokens

Gitksan
1k tokens

Lezgi
9k tokens

Natugu
12k tokens

Nyangbo
11k tokens

Tsez
47k tokens

Uspanteko
45k tokens

2023 SIGMORPHON Shared Task

Teams

COATES	LSTM Encoder-Decoder
LISNTeam	Hybrid CRF-Neural
SigMoreFun	Multilingual Pretrained Transformers
TeamSiggyMorph	BiLSTM, ByT5
Tü-CL	Straight-through gradient estimation, hard attention

2023 SIGMORPHON Shared Task

Results

MORPHEME-LEVEL ACCURACY

Submission	Arp	Ddo	Git	Lez	Ntu	Nyb	Usp	AVG	Complete?
TÜ-CL$_2$	**78.47**	**73.95**	**11.72**	**62.10**	56.32	85.24	**70.05**	62.55	**YES**
TÜ-CL$_1$	76.56	70.29	9.26	62.03	**56.38**	**86.74**	60.42	60.24	**YES**
TEAMSIGGYMORPH$_1$	-	53.19	-	28.13	31.86	66.25	59.73	47.83	
COATES$_1$	45.42	64.43	9.84	40.74	37.55	72.82	56.02	46.69	**YES**
BASELINE	44.19	51.23	8.54	41.62	18.17	14.22	57.24	33.60	**YES**

Closed Track

2023 SIGMORPHON Shared Task

Results

Morpheme-Level Accuracy

Submission	Arp	Ddo	Git	Lez	Ntu	Nyb	Usp	AVG	Complete?
TÜ-CL$_2$	**91.37**	**92.01**	50.22	**87.61**	92.32	**91.40**	**84.51**	84.21	**YES**
SIGMOREFUN$_2$	89.34	88.15	**52.39**	82.36	85.53	89.49	83.08	81.48	**YES**
LISNTEAM$_1$	-	91.39	50.80	87.17	92.60	-	82.42	80.88	
TEAMSIGGYMORPH$_2$	-	88.36	47.76	86.59	92.10	82.74	82.22	79.96	
SIGMOREFUN$_1$	91.36	84.35	47.47	80.17	88.35	85.84	80.08	79.66	**YES**
TÜ-CL$_1$	90.93	91.16	17.08	83.45	90.17	89.96	83.45	78.03	**YES**
LISNTEAM$_2$	-	-	51.09	86.52	**92.77**	-	-	76.79	
BASELINE	91.11	85.34	25.33	51.82	49.03	88.71	82.48	67.69	**YES**
SIGMOREFUN$_4$	80.81	78.24	12.74	50.00	63.39	85.30	73.25	63.39	**YES**
SIGMOREFUN$_3$	72.10	57.93	2.60	26.24	35.62	70.01	67.73	47.46	**YES**

Open Track

2023 SIGMORPHON Shared Task
Observations

- Gradient estimation for hard attention (Tü-CL; Girrback, 2023) is highly effective at the joint segmentation and glossing task
  - Also provides an interpretable model
- Multilingual training (SigMoreFun; He et al., 2023) can provide benefits to low-resource languages

What challenges remain with automated IGT systems?

Robust Generalization

Robust Generalization Strategies for Morpheme Glossing in an Endangered Language Documentation Setting. Ginn and Palmer, 2023.

Robust Generalization

- IGT corpora are often the product of a single documentation project
- Represent a limited domain of text (genre, speaker, etc)
- IGT models must **generalize** well to unseen texts for future documentation projects

Robust Generalization

We **evaluate generalization** by
splitting our dataset by **text genre**

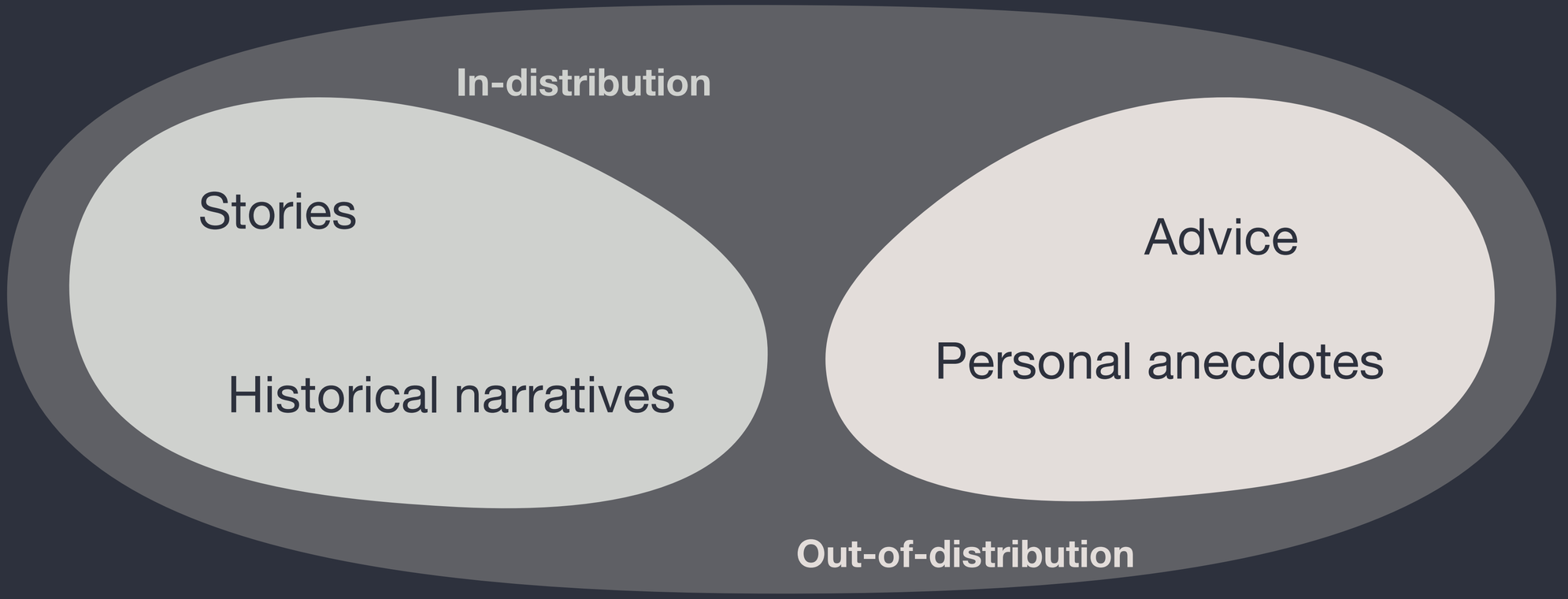
Uspanteko corpus from Palmer et al. (2009)

12k lines

29 docs

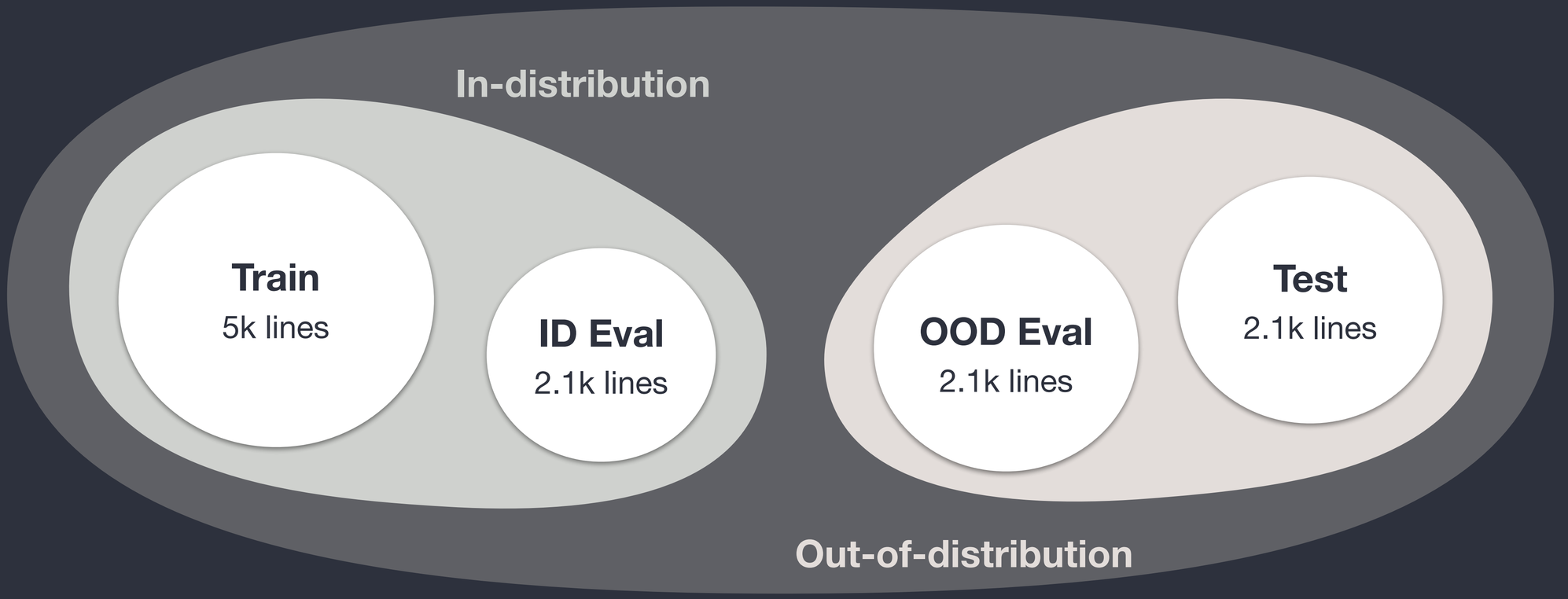
Robust Generalization

We **evaluate generalization** by
splitting our dataset by **text genre**



Robust Generalization

ID data is used for **training** and **eval**,
and OOD is used for **eval** and **testing**



Robust Generalization

Perplexity: **77.8**

Accuracy: **84.5**

Perplexity: **94.0**

Accuracy: **74.6**

We demonstrate that the OOD data performs worse for **language modeling** and **gloss generation**.

Evaluating generalization is critical for robust IGT systems that can be used in **documentation projects**.

Generalization Strategies

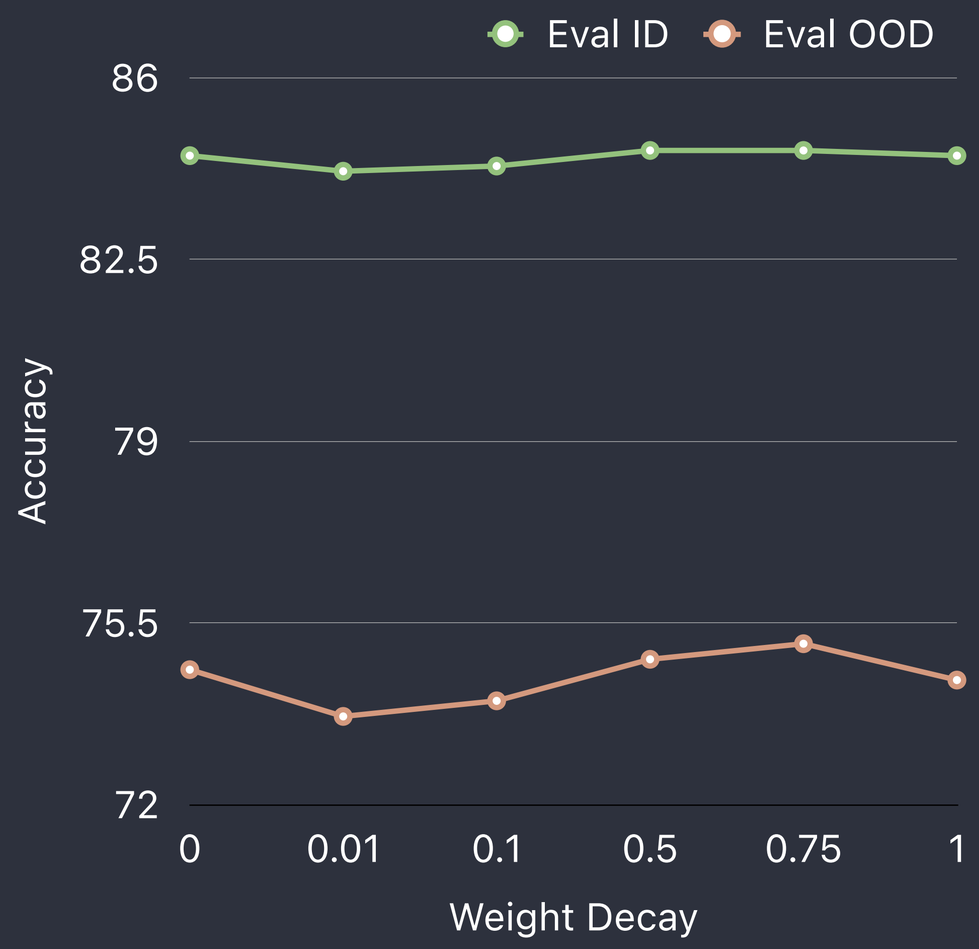
Generalization Strategies

Weight Decay

Masked Language Modeling for OOV Tokens

Iterative Pseudo-Labeling

Weight Decay



Higher weight decay helps **regularization** and **avoiding overfitting**.

Generalization Strategies

Weight Decay

Masked Language Modeling for OOV Tokens

Iterative Pseudo-Labeling

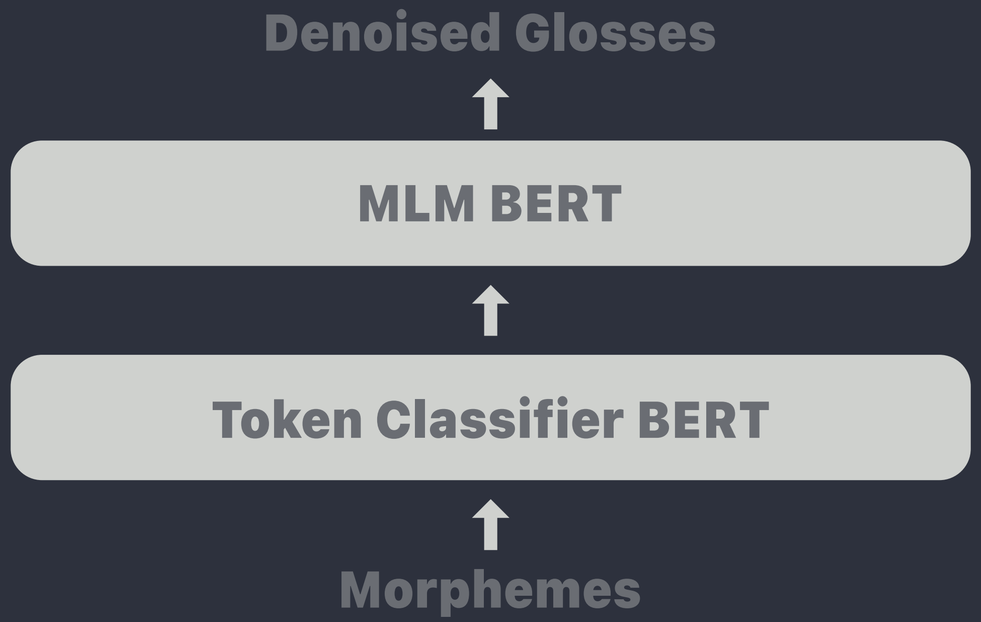
Masked Language Modeling for OOV Tokens

- Out-of-vocabulary tokens are a greater cause of error in OOD texts
  - OOD: 6.2% vs ID: 3.0%
- Transformer token classification model can make decent predictions, may be misled
- We can often recover gloss from context

Masked Language Modeling for OOV Tokens

We train a **masked language model** on gloss sequences and **apply it to the output** of the token classifier.

We achieve **limited improvement** (0.2%)



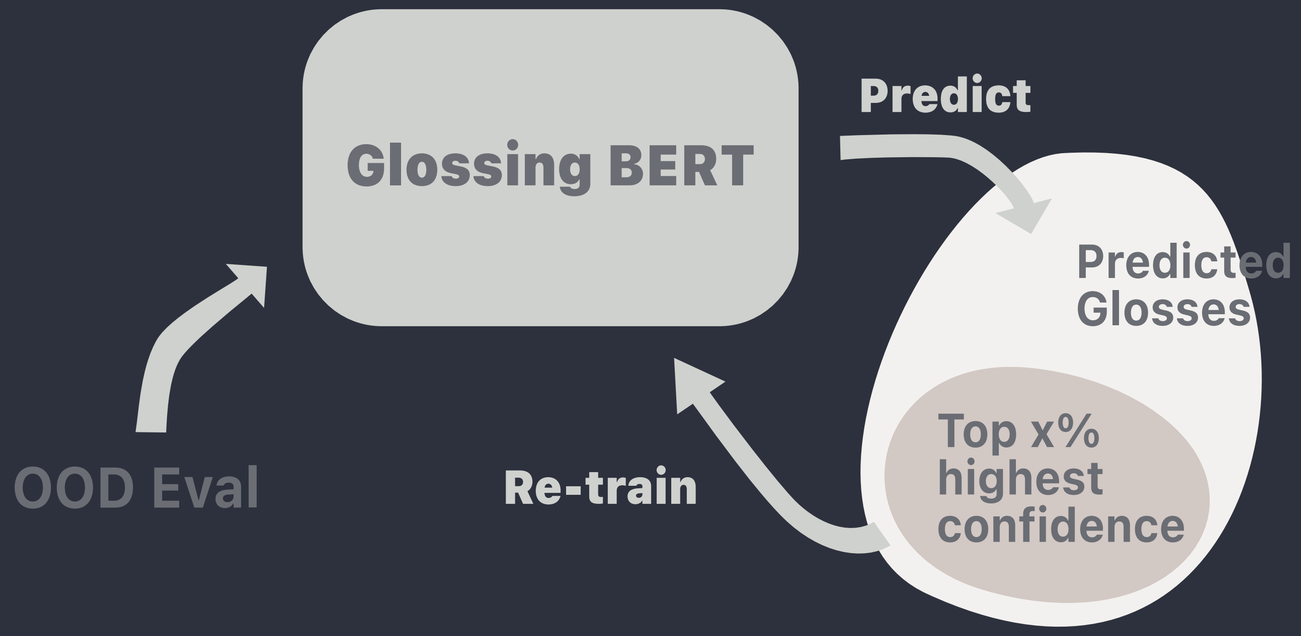
Generalization Strategies

Weight Decay

Masked Language Modeling for OOV Tokens

Iterative Pseudo-Labeling

Iterative Pseudo-Labeling



Use glossing model to do **inference on OOD data**

Select **top x% of predictions** by confidence and add to **training set**

Repeat!

Results


Test OOD
Eval OOD
78
77
76
75
74
Baseline
+ Weight Decay Optimization
+ Denoising
+ Iterative Pseudo-Labeling

Discussion

- Training strategies can improve robustness a limited amount
- Distributional shift remains a difficult problem for IGT models

Human-Like Glossing

Taxonomic Loss for Morphological Glossing of Low-Resource Languages.
Ginn and Palmer, 2023. Preprint.

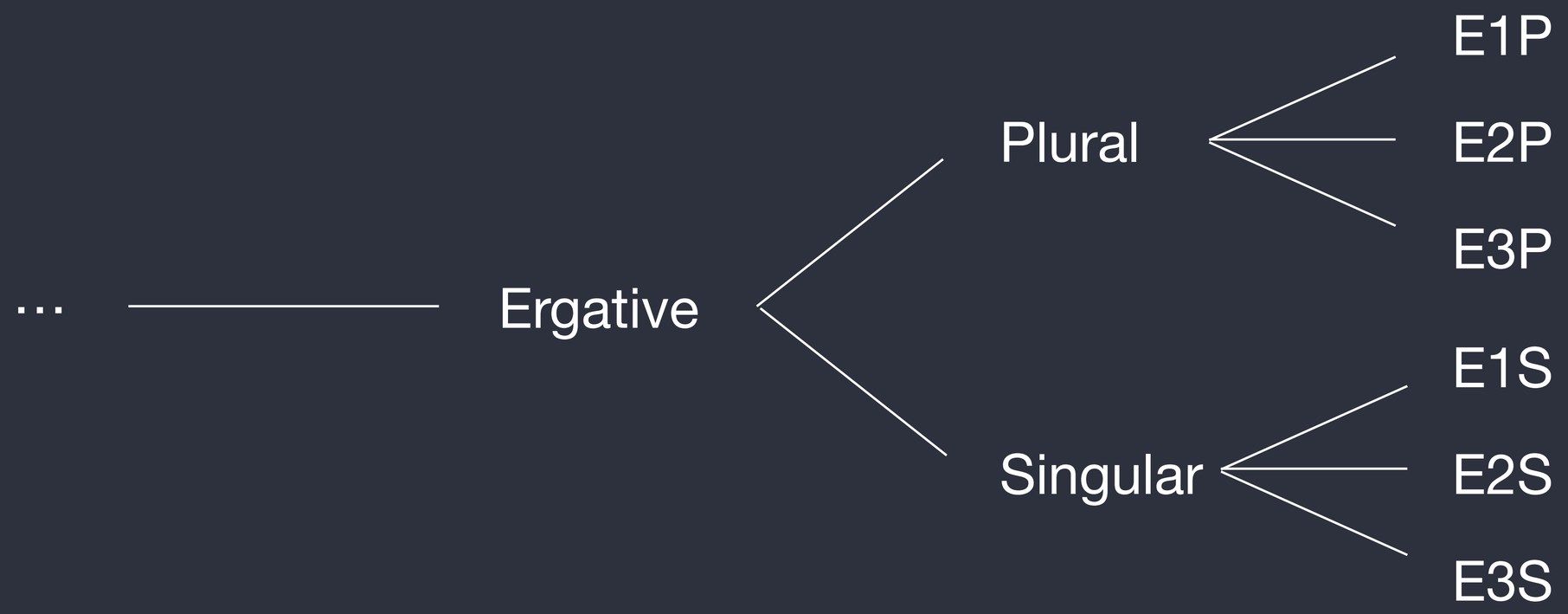
Human-Like Glossing

Motivation

- In a human-in-the-loop documentation setting, models should suggest glosses that are **accurate** but also **predictable**
- Currently, SOTA glossing models achieve high accuracy but make **unexpected** and **unintuitive** **errors**
- Meanwhile, Human annotators may consider **sets of similar morphemes** when choosing the appropriate label

Human-Like Glossing

Taxonomy of Morphemes



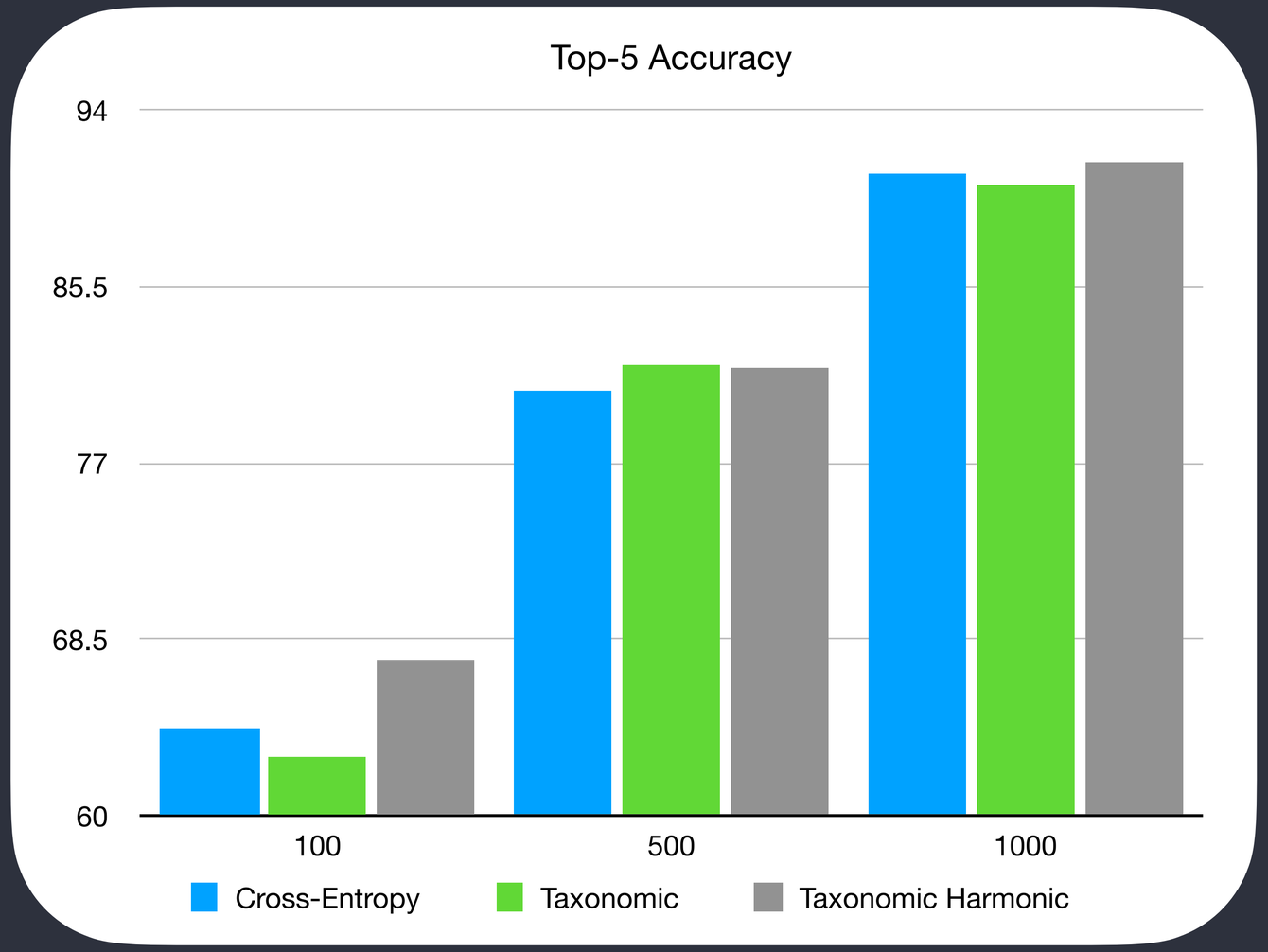
Human-Like Glossing

Taxonomic Loss Function

- We take inspiration from computer vision in natural sciences (plant species: Wu et al., 2019; seafloor images: Nourani-Vatani et al., 2015) which also have natural taxonomies of labels
- **Taxonomic loss function** penalizes being further away from correct label in taxonomy
  - At each level of taxonomy, sum logits and apply cross-entropy loss)
  - Sum weighted losses across levels

Human-Like Glossing

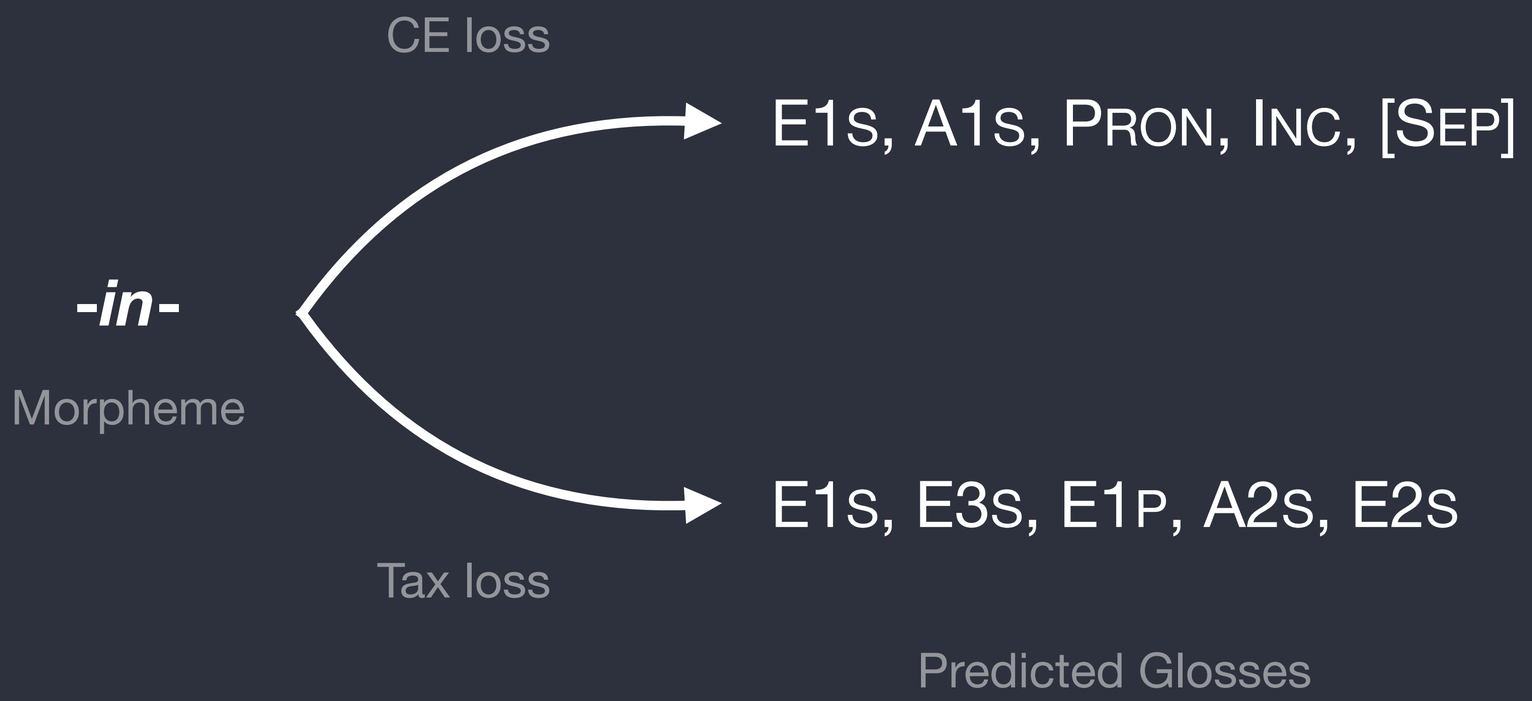
Results



Taxonomic loss functions may help models suggest correct label

Human-Like Glossing

Results



Human-Like Glossing
Discussion

- Taxonomic loss may have small benefit to performance
- Models make more intuitive predictions, don’t confuse annotators
- Can predict glosses which are rarely (or even never) observed

Overall Takeaways

- Automated IGT Glossing models are becoming more capable with techniques such as multilingual transfer and hard attention
- IGT models must be robust to distributional shift for real-world usage
- IGT models can be steered to make human-like predictions

Thank you!

This material is based upon work supported by the National Science Foundation under Grant No. 2149404, “CAREER: From One Language to Another”. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.