ÚFAL Submission for SIGTYP Supervised Cognate Detection Task

Tomasz Limisiewicz

Task Definition

- **Cognates** are pairs with similar meanings and come from the same root in an ancestral language.

  de: 'Vater' <-> en: 'father'

- Multilingual **derivatives** are words borrowed from another language potentially with some modification.

  fr: 'restaurant' -> en: 'restaurant'

"A Dictionary of Linguistics and Phonetics" David Crystal, 2008

Data

The training data contained over 232,482 pairs of bilingual words together with relationship labels (cognate/derivative/no relation) from Wiktionary

The dataset covered 34 European languages

The task was evaluated on 876 test pairs

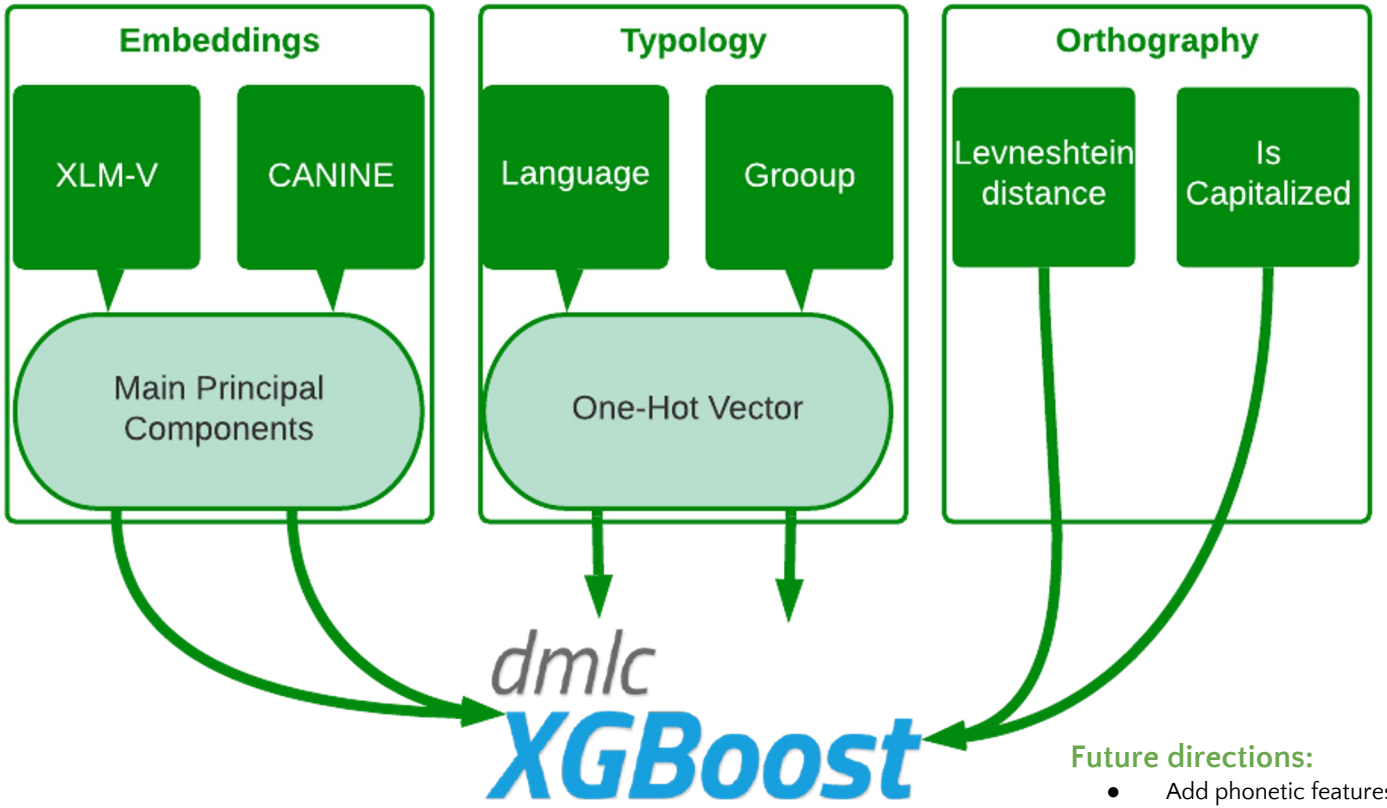
For validation I sampled 10% of training dataset

Gradient Boosted Trees

- Method ensembling predictions of large set of decision trees (with gradient search).
- Strong performance for discrete and real features.
- Feature importance analysis can help interpret predictions.

Parameter	Search Range	Selected
eta	0.01 - 0.3	0.275
gamma	0 - 5.0	0.642
maximum depth	3 - 20	12
minimum child weight	1 - 6	4
subsample	0.6 - 1.0	0.723
column sample (tree)	0.6 - 1.0	0.919
column sample (node)	0.6 - 1.0	0.749
column sample (level)	0.6 - 1.0	0.998
lambda	0 - 5.0	1.507
alpha	0 - 5.0	1.138

Features Selection



Future directions:

- Add phonetic features
- Extend typological genealogical analysis

Results & Accumulation

- Final results **R 86% P 89% F1 87%**

	Features	Train Acc	Train F1	Validation Acc	Validation F1
1	Language ID	75.9	64.2	76.1	64.2
2	(1) + Group ID	76.6	64.7	76.7	64.6
3	(2) + Capitalized	78.4	66.3	78.6	66.4
4	(3) + Levenshtein	83.1	70.6	83.0	69.8
5	(3) + Embeddings	97.2	94.2	92.6	80.3
6	(4) + Embeddings No weighting	**98.4**	95.8	**93.8**	79.6
7	**(4) + Embeddings**	97.8	**95.3**	93.7	**82.7**

Results & Accumulation

- Final results **R 86% P 89% F1 87%**
- Good results for language ID alone

	Features	Train Acc	Train F1	Validation Acc	Validation F1
1	Language ID	75.9	64.2	76.1	64.2
2	1 + Group ID	76.6	64.7	76.7	64.6
3	2 + Capitalized	78.4	66.3	78.6	66.4
4	3 + Levenshtein	83.1	70.6	83.0	69.8
5	3 + Embeddings	97.2	94.2	92.6	80.3
6	4 + Embeddings No weighting	**98.4**	**95.8**	**93.8**	79.6
7	**4 + Embeddings**	97.8	95.3	93.7	**82.7**

Results & Accumulation

- Final results **R 86% P 89% F1 87%**
- Good results for language ID alone
- Character-level signal and embeddings important for improving results

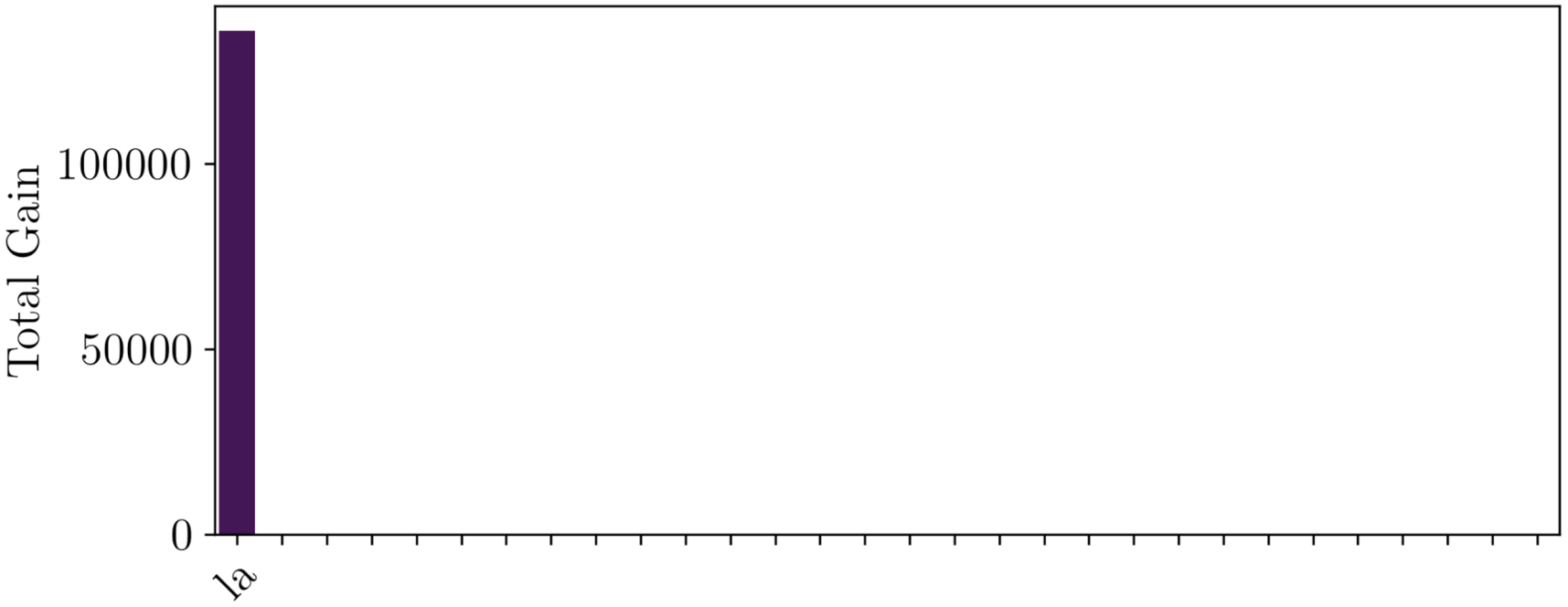
	Features	Train Acc	Train F1	Validation Acc	Validation F1
1	Language ID	75.9	64.2	76.1	64.2
2	(1) + Group ID	76.6	64.7	76.7	64.6
3	(2) + Capitalized	78.4	66.3	78.6	66.4
4	(3) + Levenshtein	83.1	70.6	83.0	69.8
5	(3) + Embeddings	97.2	94.2	92.6	80.3
6	(4) + Embeddings No weighting	**98.4**	**95.8**	**93.8**	79.6
7	**(4) + Embeddings**	97.8	95.3	93.7	**82.7**

Results & Accumulation

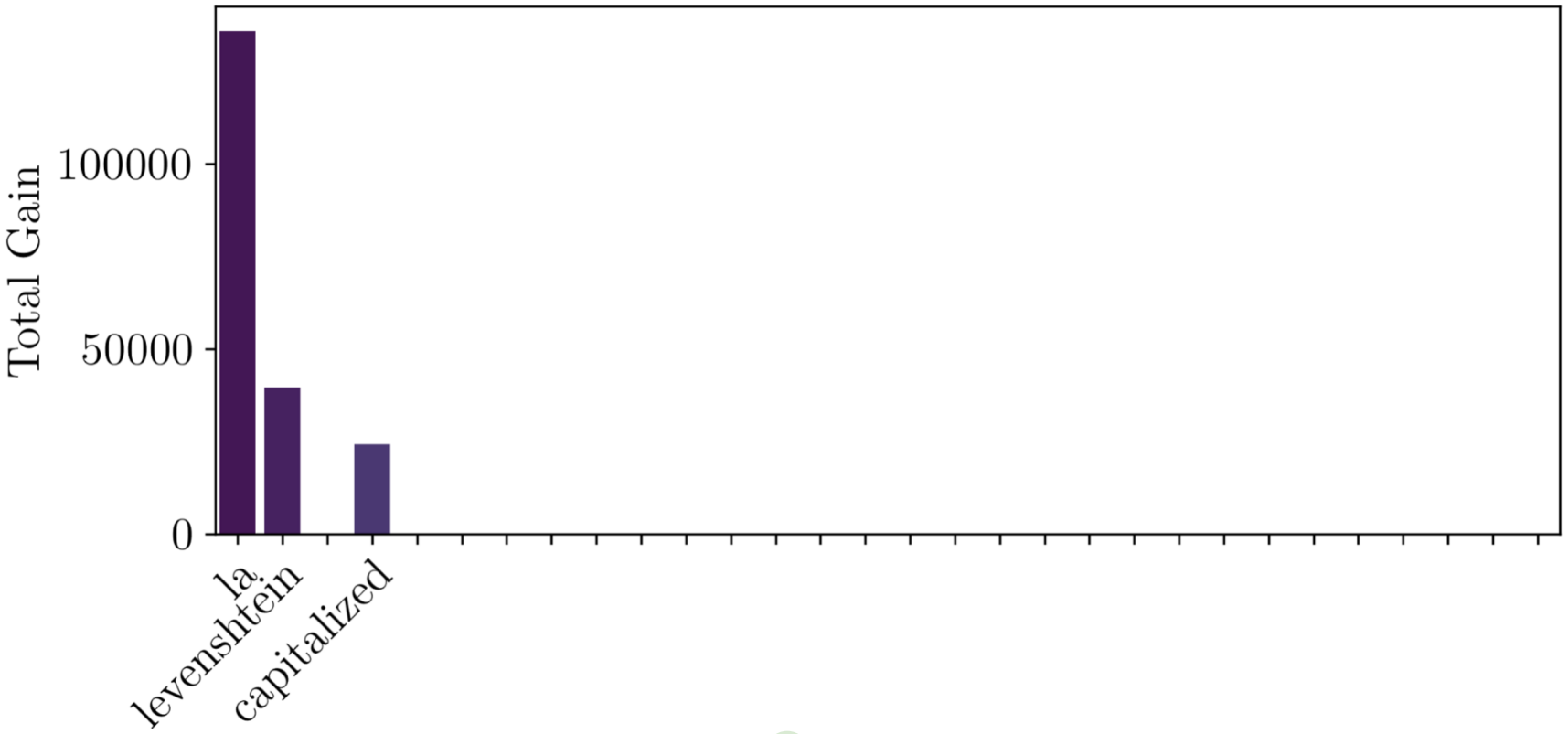
- Final results **R 86% P 89% F1 87%**
- Good results for language ID alone
- Character-level signal and embeddings important for improving results
- Weighteing counters unbalance in training data and improves F1

	Features	Train Acc	Train F1	Validation Acc	Validation F1
1	Language ID	75.9	64.2	76.1	64.2
2	(1) + Group ID	76.6	64.7	76.7	64.6
3	(2) + Capitalized	78.4	66.3	78.6	66.4
4	(3) + Levenshtein	83.1	70.6	83.0	69.8
5	(3) + Embeddings	97.2	94.2	92.6	80.3
6	(4) + Embeddings No weighting	**98.4**	**95.8**	**93.8**	79.6
7	(4) + **Embeddings**	97.8	95.3	93.7	**82.7**

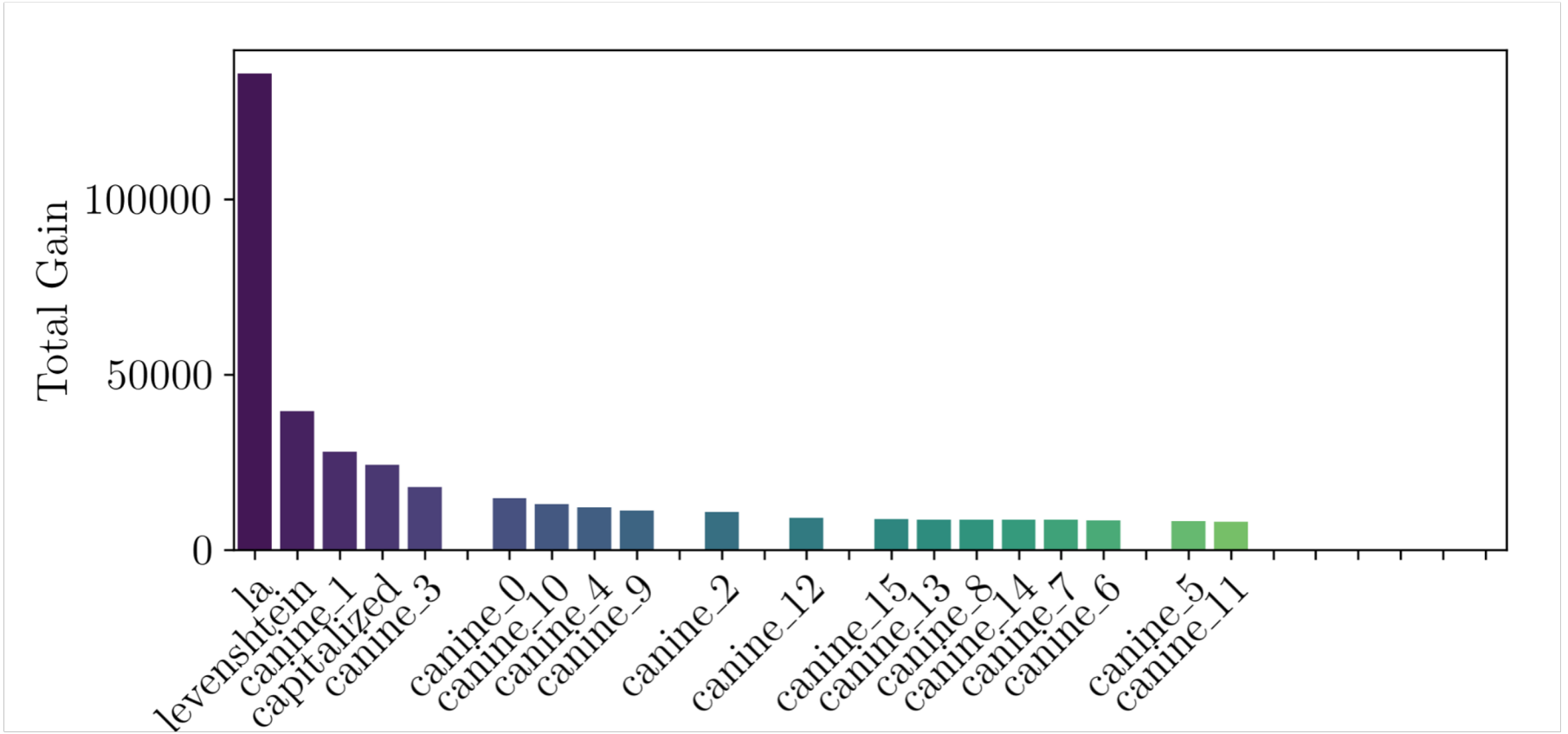
Feature Importance



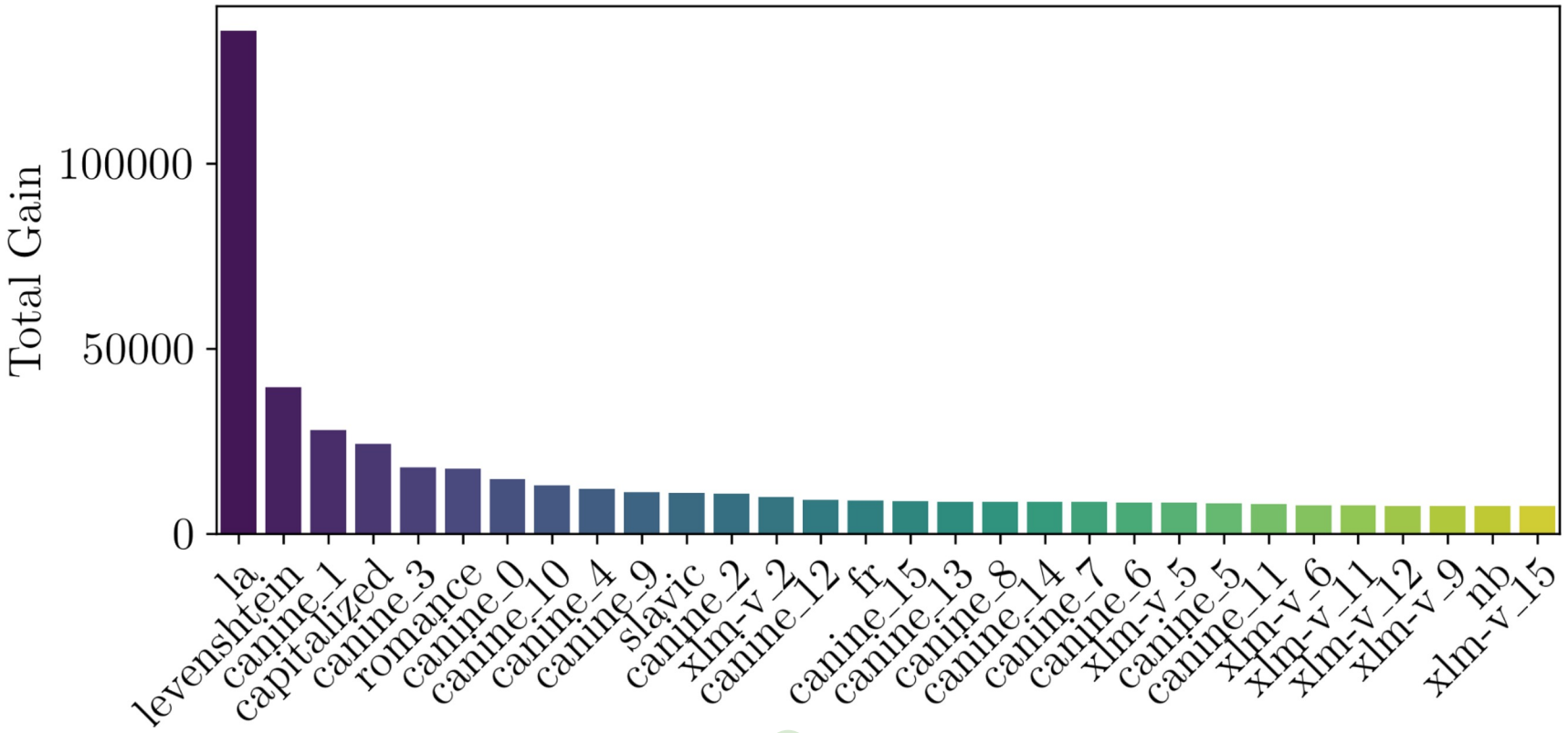
Feature Importance



Feature Importance



Feature Importance



Summary:

1. Language identity is a strong predictor
2. Character-level signals are important for cognate identification
3. Extending the proposed method is possible and encouraged

Thank You
For your Attention!