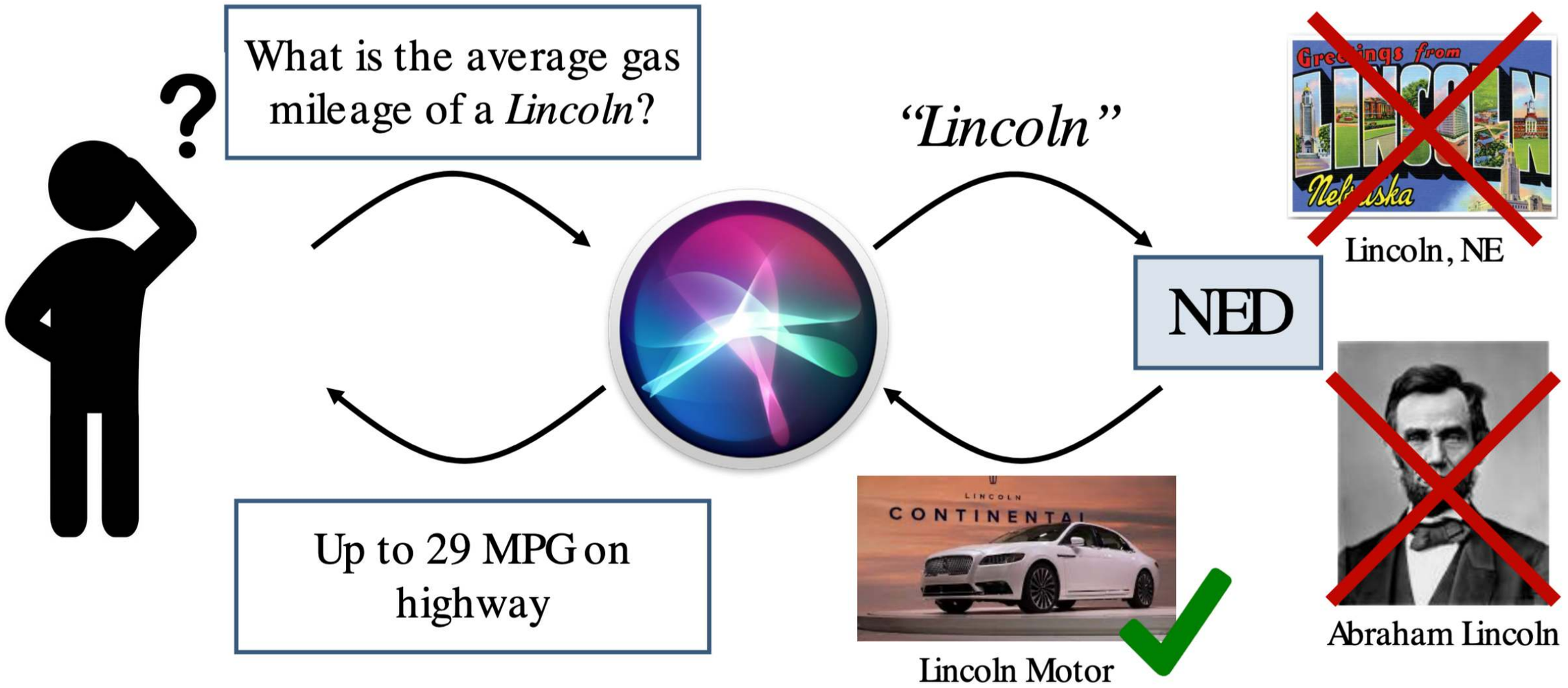
Minimalist Entity Disambiguation for Mid-Resource Languages

Benno Kruit, Vrije Universiteit Amsterdam

SustaiNLP 2023

Named Entity Disambiguation



from: http://ai.stanford.edu/blog/bootleg/

Motivation

Why Minimalist Disambiguation?

- NED models are often **very large**
- Examples:
  - **Spotlight** (statistical, multilingual)
    up to 2 GB (English model)
  - **mGENRE** (neural, multilingual)
    11 GB
  - **BootLeg** (neural, SoTA)
    5 GB
- *Compressed sizes*! Larger in practice

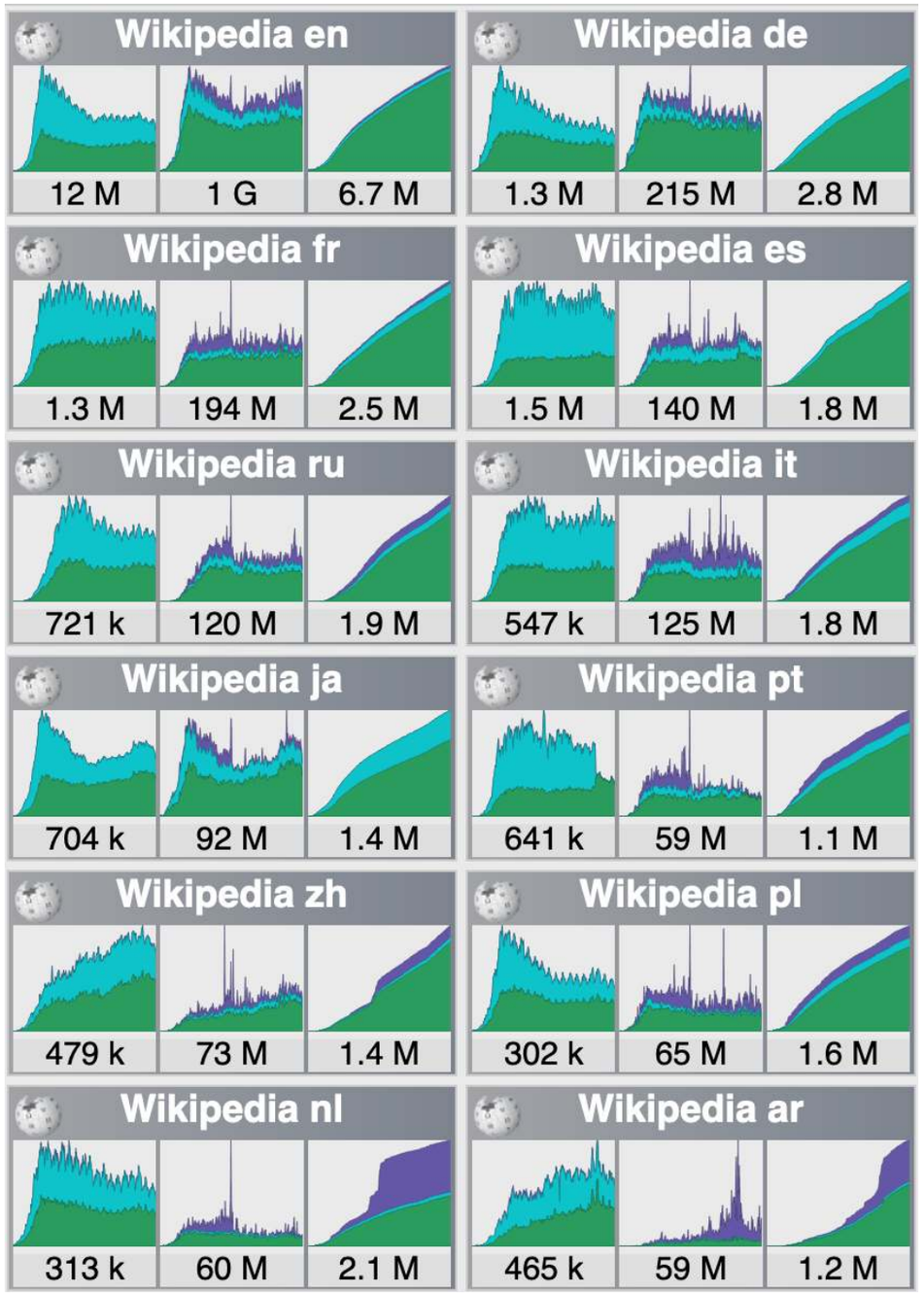
BOOTLEG

VU VRIJE UNIVERSITEIT AMSTERDAM

Motivation

Why Minimalist Disambiguation?

- **Tune** models per use-case & language
  - Aim for distribution **head** or **tail**?
    - Most important **domains**?
    - Simplifying **assumptions**?
- Training data determines strategy
  - Data **Size** & **Quality**
- Small models:
  flexible & **sustainable**



Users / edits / articles

Observations
on benchmark data

- **Mewsli-9**: articles from WikiNews
  - **58,717 non-parallel articles** from 2010-2019 (written by volunteers)
  - **Automatic annotations** hyperlinks to Wikipedia (any language)
  - **9 Languages** Japanese, German, Spanish, Arabic, Serbian, Turkish, Persian, Tamil, ~~English~~ **Dutch** (due to focus on *mid-resource* languages)

Observations
on benchmark data

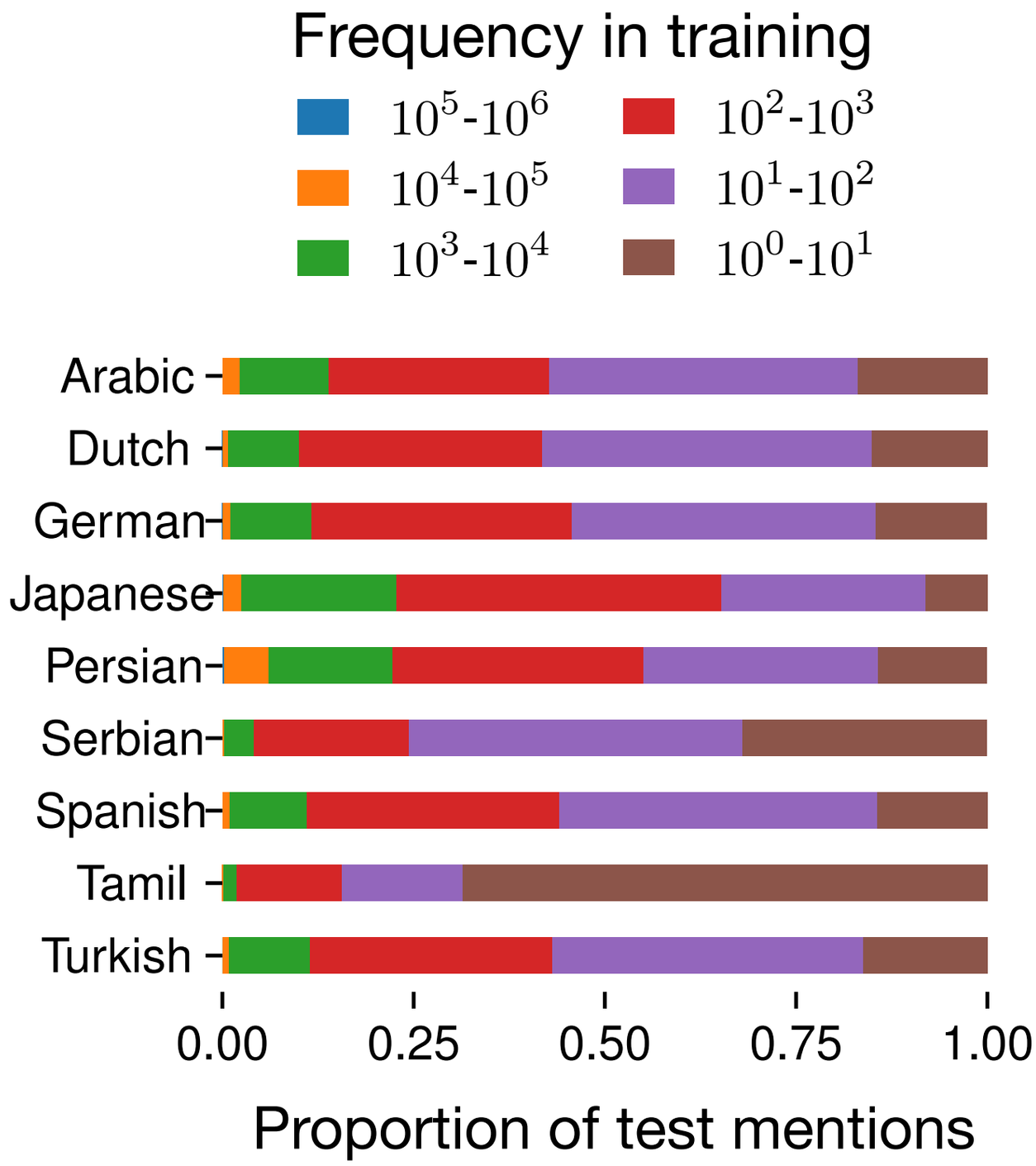
- **Train** = Wikipedia
  Test = Mewsli-9

- “mention” = name-entity pair

- How often do mentions in the *test* data occur in *training*?
  - Mostly **10 - 10.000 times**

- Long tail has lots of data

- **Mid-resource:**
  enough data to train on!

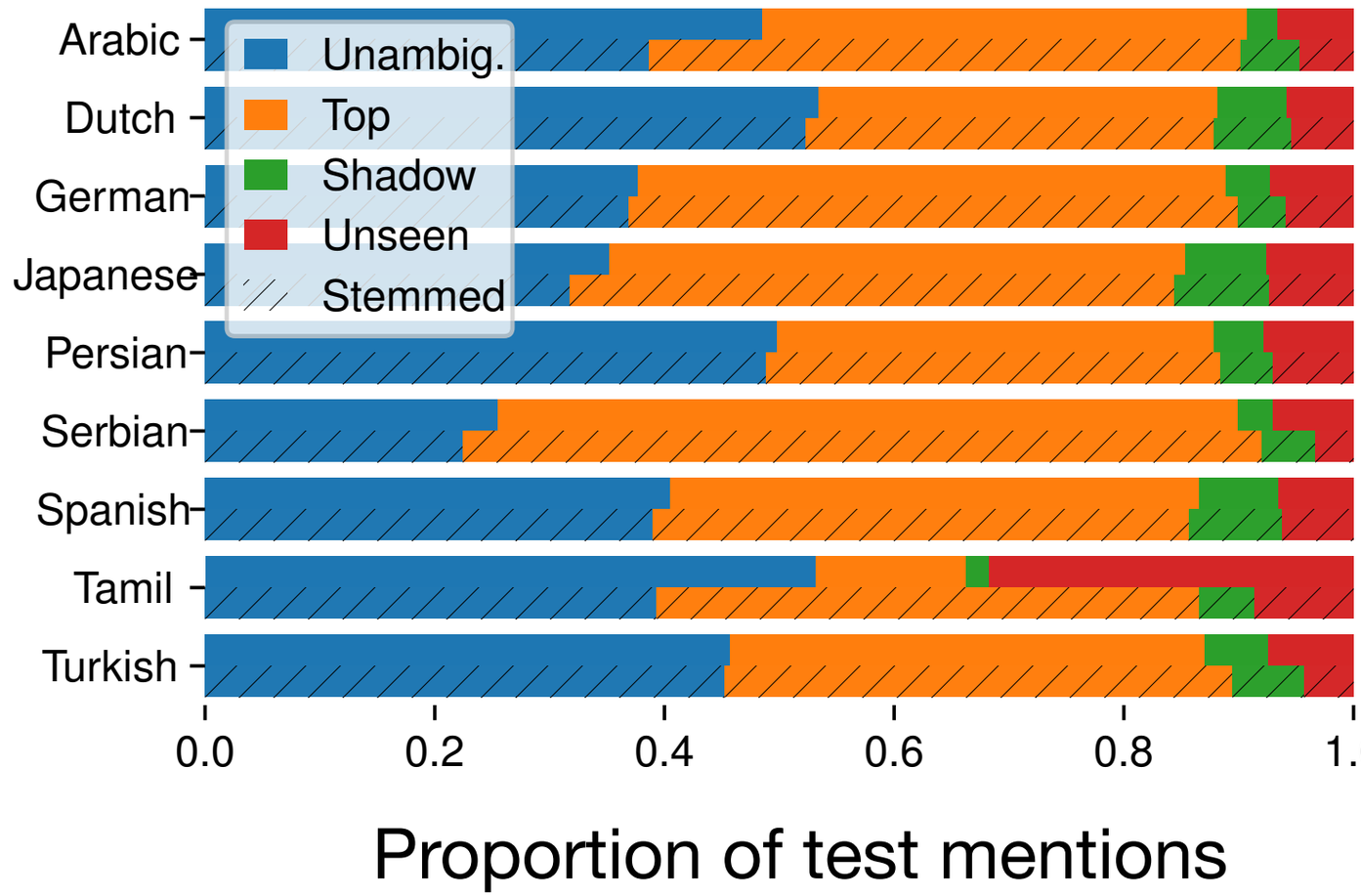


Observations
on benchmark data

- **Training rank** of test mention
  - e.g. “london” →
    1. **London_(UK)**
    2. **London_(Canada)**
    (etc)
    unseen: **Jane_London**

- Narrow bound of **shadowed** mentions → this is the task

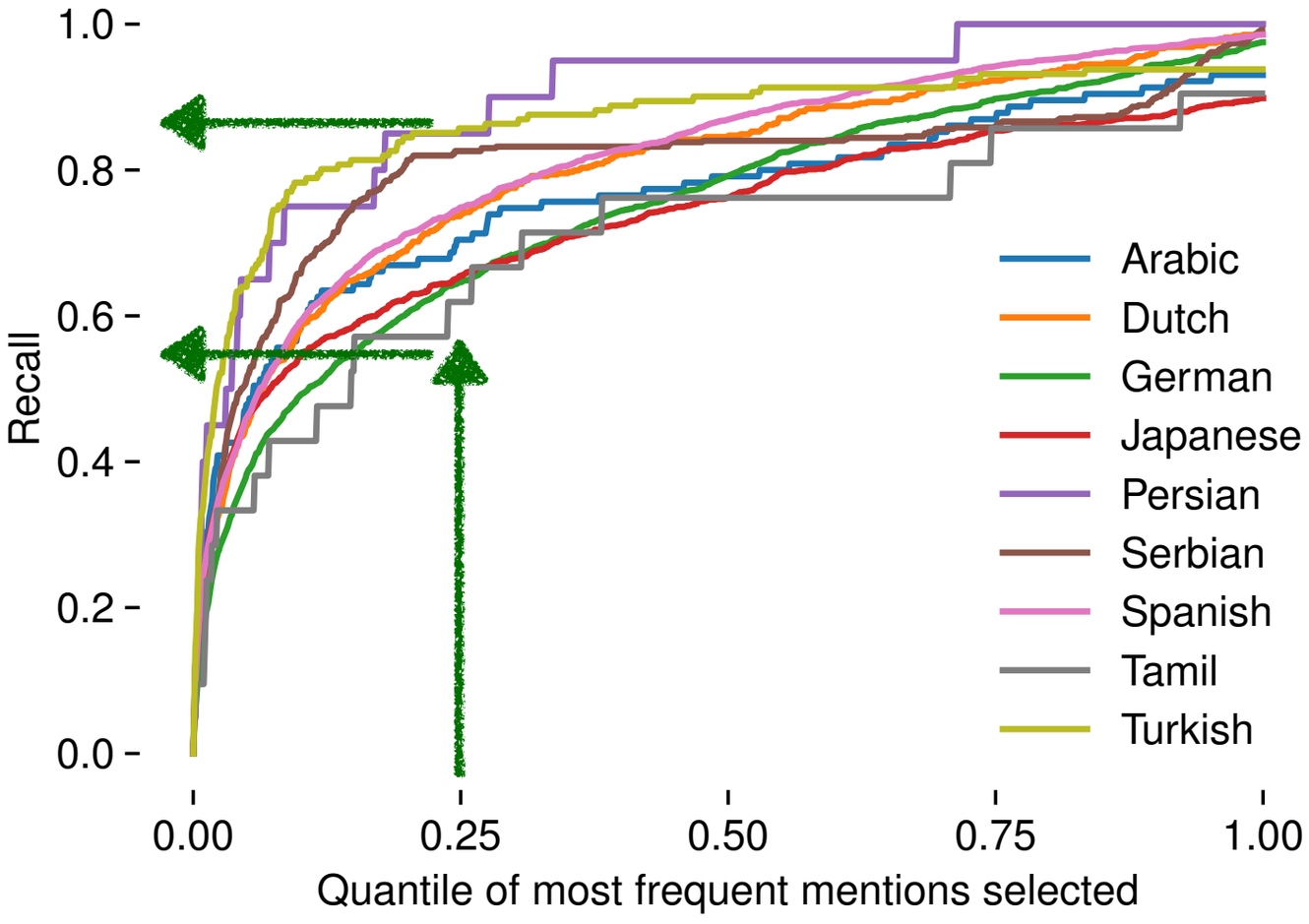
- **Stemming:** (/////)
  - Inflections differ per language
  - More **ambiguous** & more **seen**



VU VRIJE UNIVERSITEIT AMSTERDAM

Approach
for minimalist NED

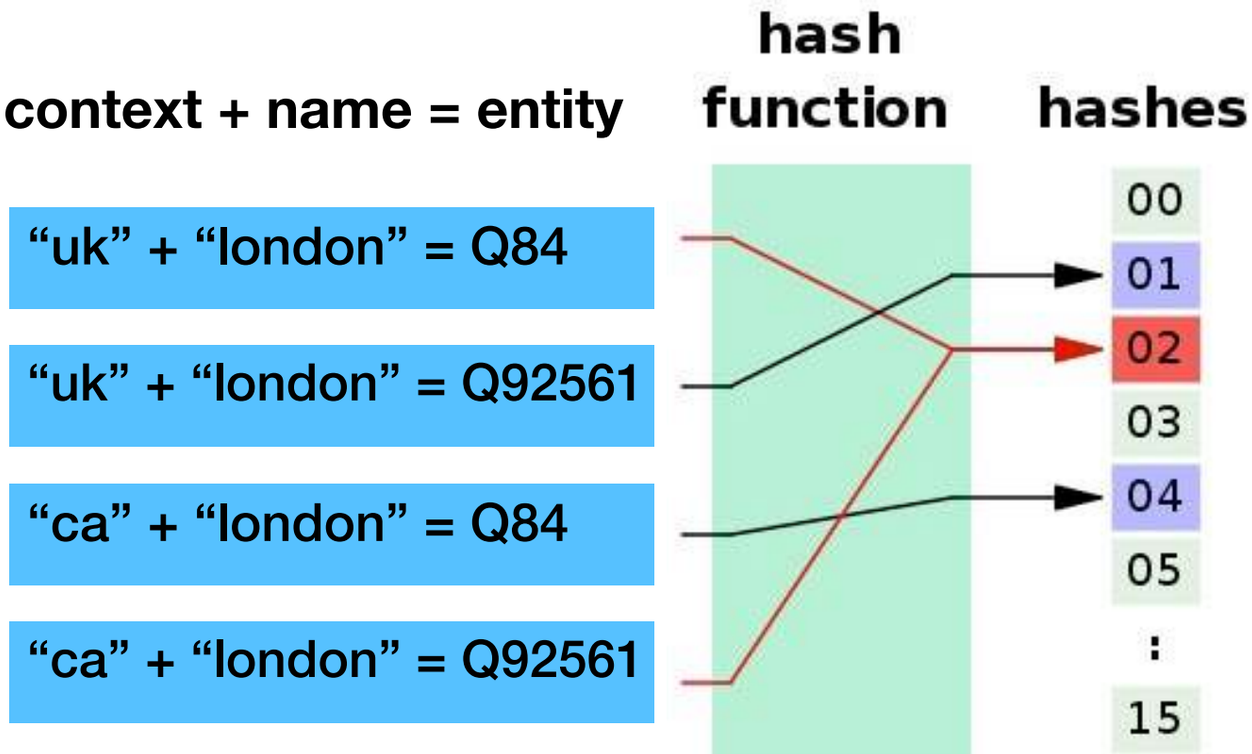
- Collect & filter candidates
    - Clean with **heuristics**
- Use top % of mentions
    - **Trade-off:**
      Simplicity vs. accuracy
    - Different per language
    - **Top 25% →**
      55-85% recall
- **Fallback** to most frequent entity



VU VRIJE UNIVERSITEIT AMSTERDAM

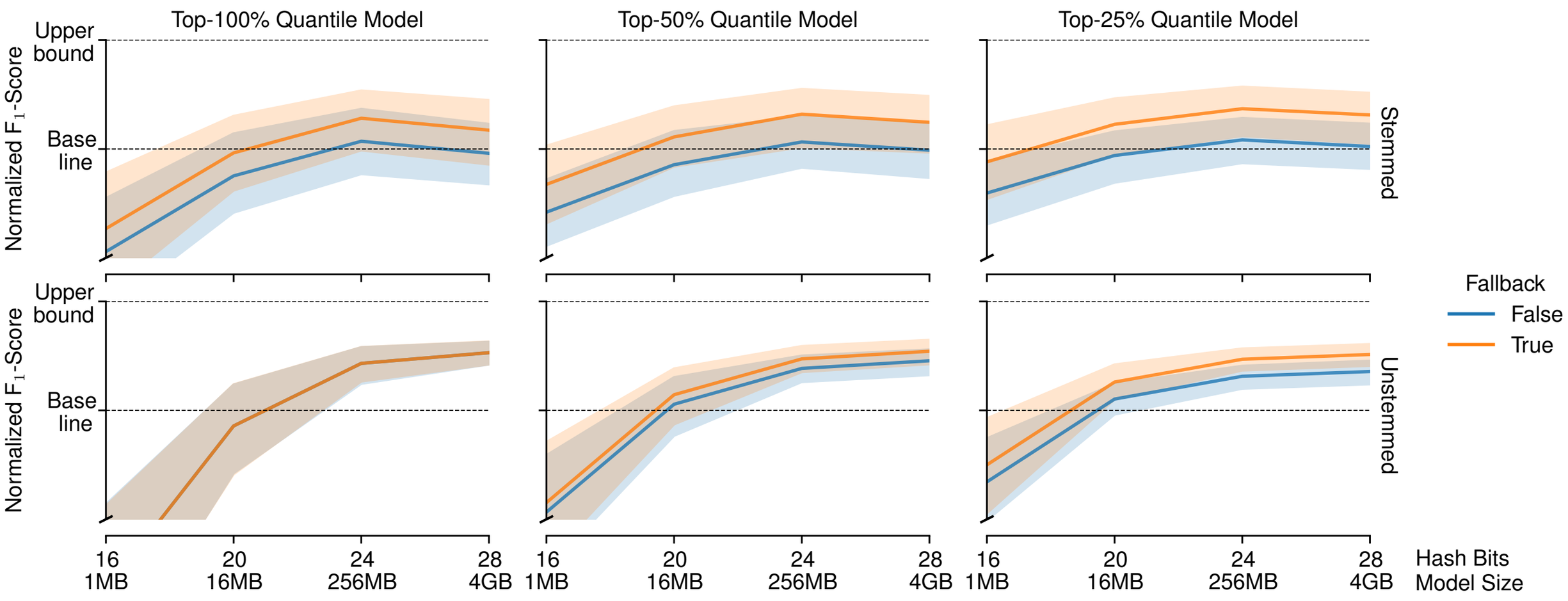
Approach
for minimalist NED

- Logistic Regression on Bag-of-Words features
- Feature Hashing: **Vowpal Wabbit**
  - **Trade-off:** Space/speed vs. accuracy
  - Collisions = regularization
  - (Features * Entities) matrix
    ↓
    Fixed size parameter space + variable collision rate
- **Hyperparameter**: Number of **Bits**



Evaluation

per setting on all languages



- Strong **diminishing returns** on model size
- Smaller models with quantile-based **candidate selection**
- **Stemming** helps some languages, but mixed effects overall

VU VRIJE UNIVERSITEIT AMSTERDAM

Evaluation
per language

	Baseline	Best Model$_{\text{bits}}$	Upper Bound
Arabic	.87	$.82_{28}/.\mathbf{89}_{28}$	.93/.91
Dutch	.63	$.77_{28}/.\mathbf{78}_{28}$	.84/.83
German	.80	$.84_{28}/.\mathbf{85}_{28}$	.90/.88
Japanese	.80	$.81_{28}/.83_{28}$	.91/.89
Persian	.85	$.\mathbf{88}_{28}/.\mathbf{88}_{24}$	.91/.90
Serbian	.76	$.83_{28}/.80_{28}$	.89/.83
Spanish	.71	$.78_{28}/.\mathbf{81}_{28}$	.89/.88
Tamil	.61	$.\mathbf{75}_{24}/.63_{24}$	.77/.64
Turkish	.80	$.80_{28}/.81_{28}$	.91/.87

F1 score (stemmed / unstemmed)

- Reasonable performance, but clear room for improvement
- Optimal parameters are **different per language**: tuning!

VU VRIJE UNIVERSITEIT AMSTERDAM

Explanation
of model parameters

Utrecht_(stad)	utrecht 1.30	stad [illegible]	provincie -1.02	schilderij 0.96	nederlands 0.95	binnenstad 0.89	museum 0.88
Utrecht_(provincie)	provincie 2.03	geografie 1.09	baarn 1.05	waterschap 1.03	gemeentelijk 0.92	wakkerendijk 0.92	provincies 0.80
Utrecht_(Zuid-Afrika)	categorie -0.59	[illegible] -0.38	rotterdammers -0.37	zuid 0.36	type 0.36	republiek 0.34	is -0.34
Universiteit_Utrecht	provincie -0.68	universiteit 0.65	universiteiten 0.62	hoogleraar [illegible]	bisschop -0.52	plaats -0.47	gemeente -0.41
FC_Utrecht	categorie -0.50	volksvertegenwoordiging -0.46	eibert -0.44	club 0.43	oormalig 0.40	fc 0.38	roelandszoon -0.37

- Useful for “data debugging”

Conclusion

Minimalist Entity Disambiguation for Mid-Resource Languages

- How much do we *need*?
  - Simple features: ~256 MB model
  - … but probably we need better features
- How much can we *leverage*?
  - Entity features?
  - Better filters / combinations?
- Language differences
  - Inflection matters

Future Work

- Robustness evaluation
- (Contextual / Hash) Word Embeddings
- Feature Selection

Code & Data

http://github.com/**bennokr/miniNED**

Benno Kruit (b.b.kruit@vu.nl)

Thank you!