Dependency Parsing & Information Extraction in Low-Resource Scenarios

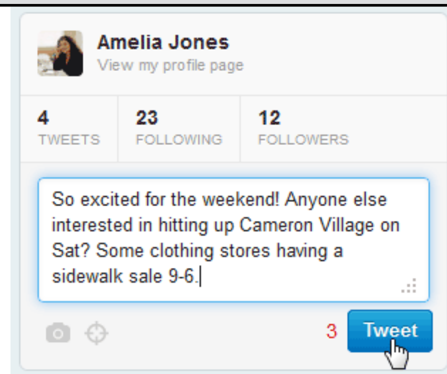
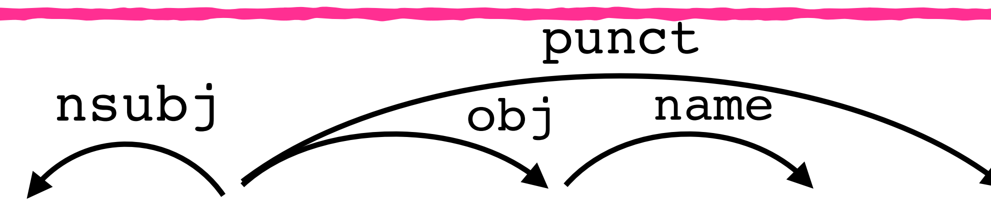
Barbara Plank
LMU Munich, Center for Information & Speech Processing (CIS)
&
IT University of Copenhagen, NLPnorth

April 20, 2022
Gothenburg (CLASP seminar)

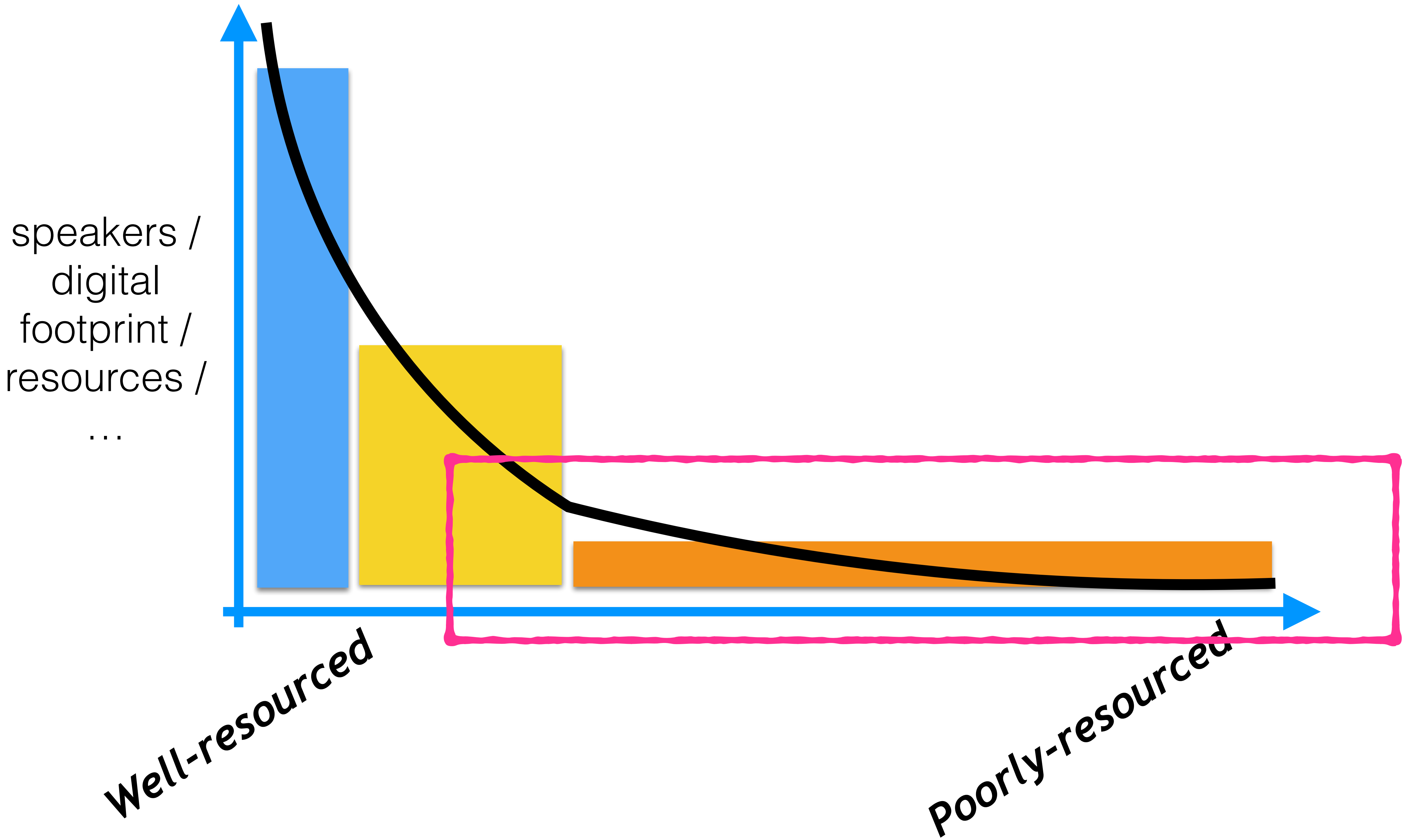
NLP Tasks: Learning from <X,Y>

➡ Time-intensive
➡ Expensive

human-annotated examples

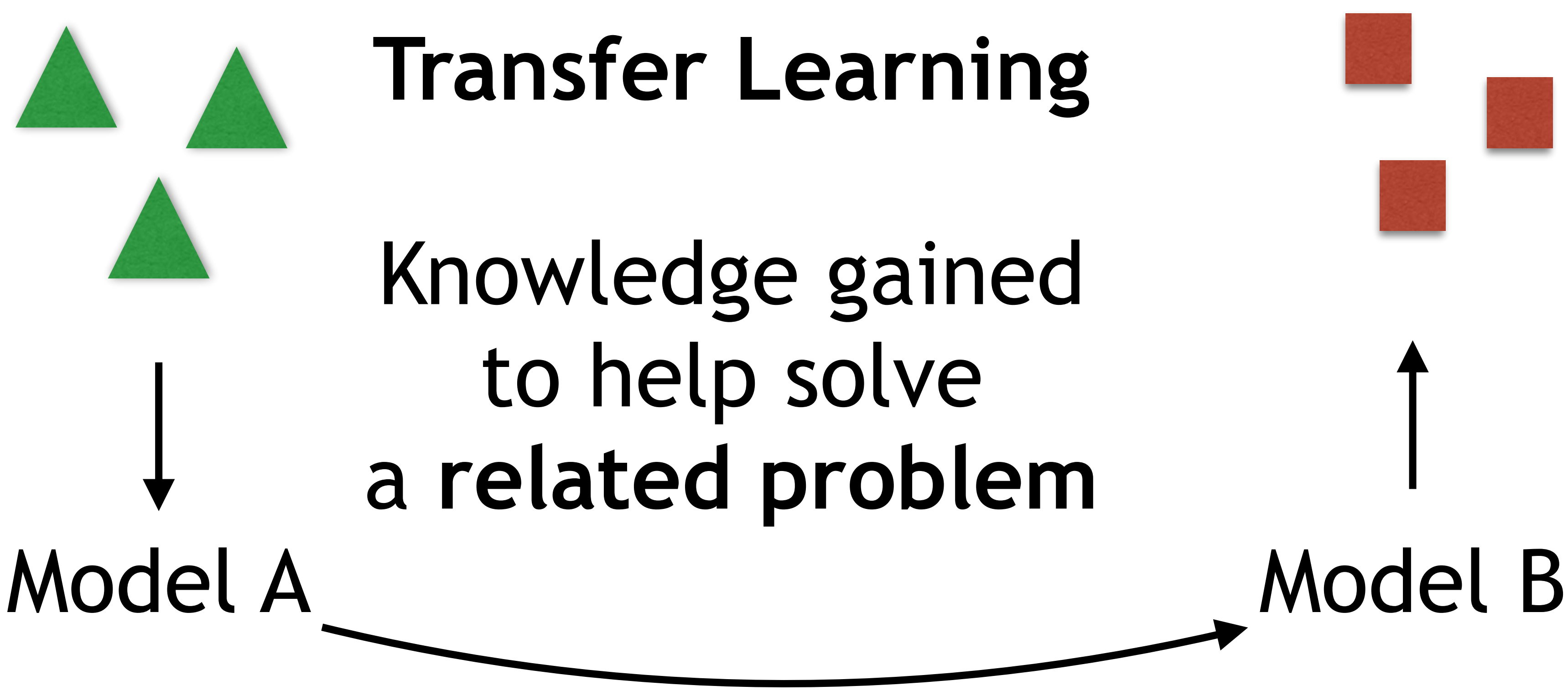
<table>
<tr><th></th><th>X (input)</th><th>Y (output)</th></tr>
<tr><td>Sentiment Analysis</td><td>
</td><td></td></tr>
<tr><td>Dependency Parsing</td><td>I like Vince Gilligan .</td><td>
</td></tr>
<tr><td>Information Extraction</td><td>Citigroup has taken over EMI,</td><td>CompanyAcquired(Citigroup, EMI)</td></tr>
</table>

Labeled data is scarce



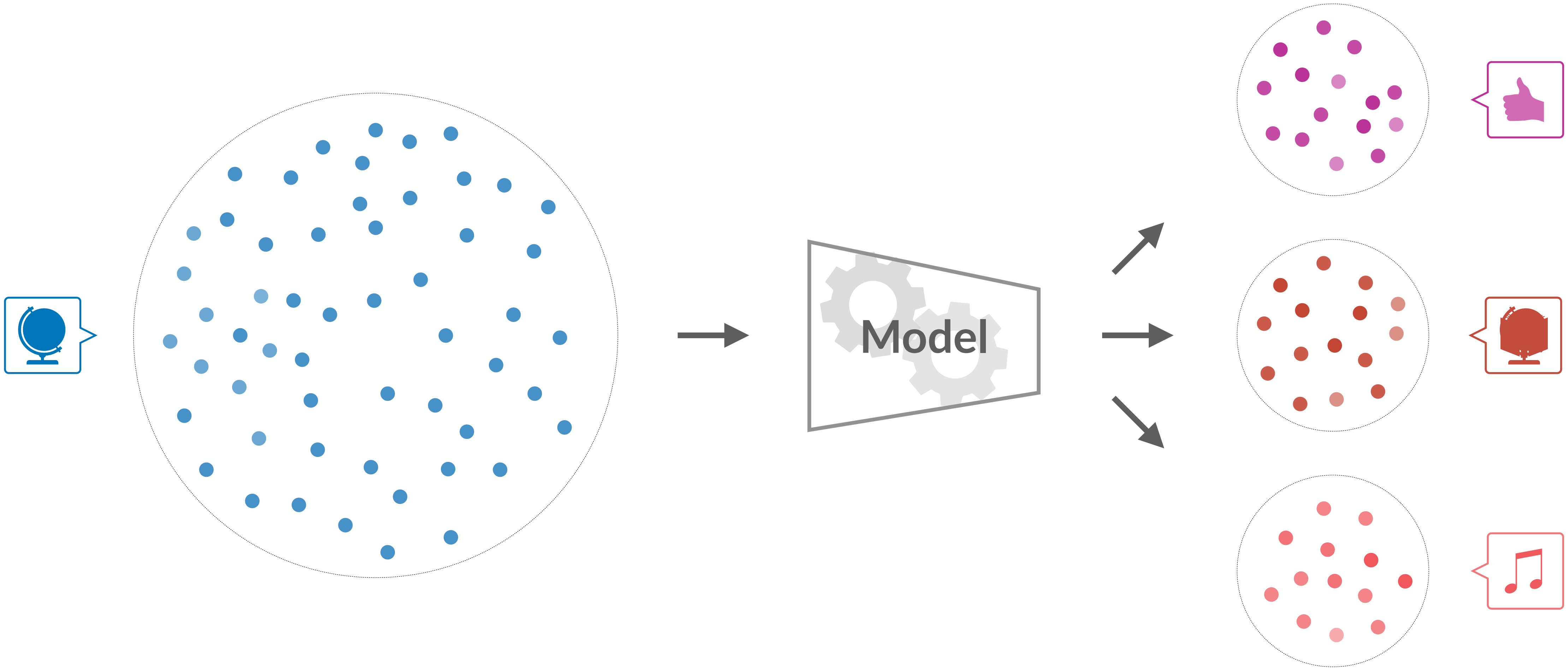
What to do about it?

Adaptation / Transfer Learning



Roadmap

1. How useful is (fortuitous) meta-data for low-res parsing?
2. How impactful are segment embeddings for low-res NLP?
3. To what extent does auxiliary data help limited training data?


Model

Genre as Weak Supervision for Cross-lingual Dependency Parsing

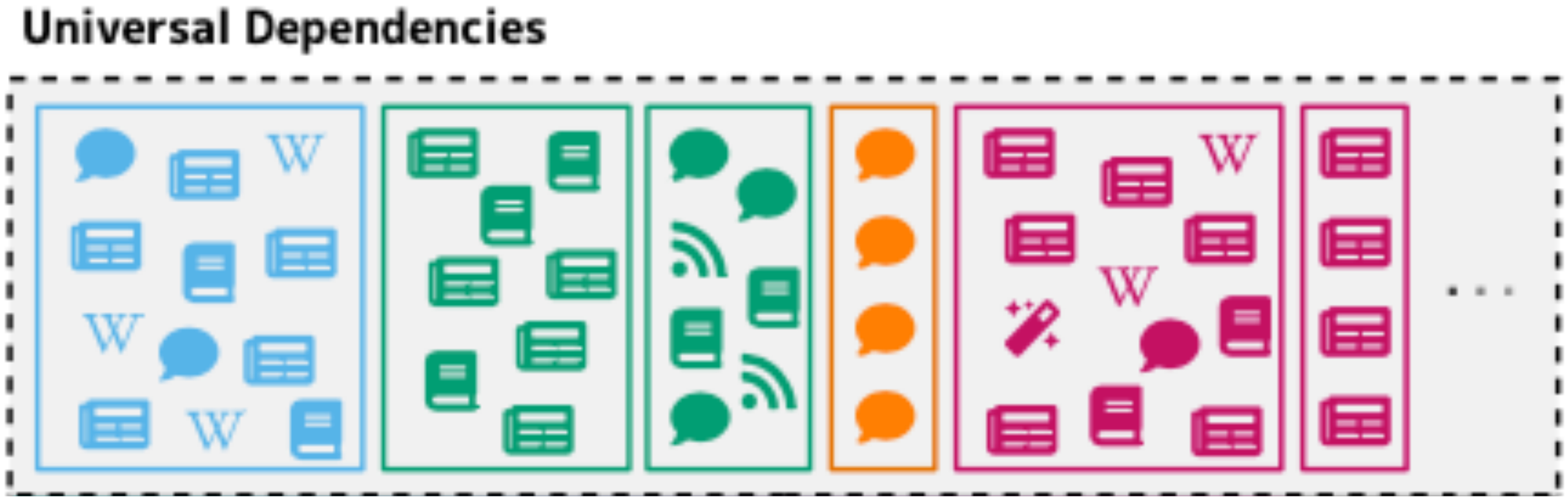
Max Müller-Eberstein and **Rob van der Goot** and **Barbara Plank**
Department of Computer Science
IT University of Copenhagen, Denmark
`mamy@itu.dk, robv@itu.dk, bapl@itu.dk`

Part 1


EMNLP, 2021

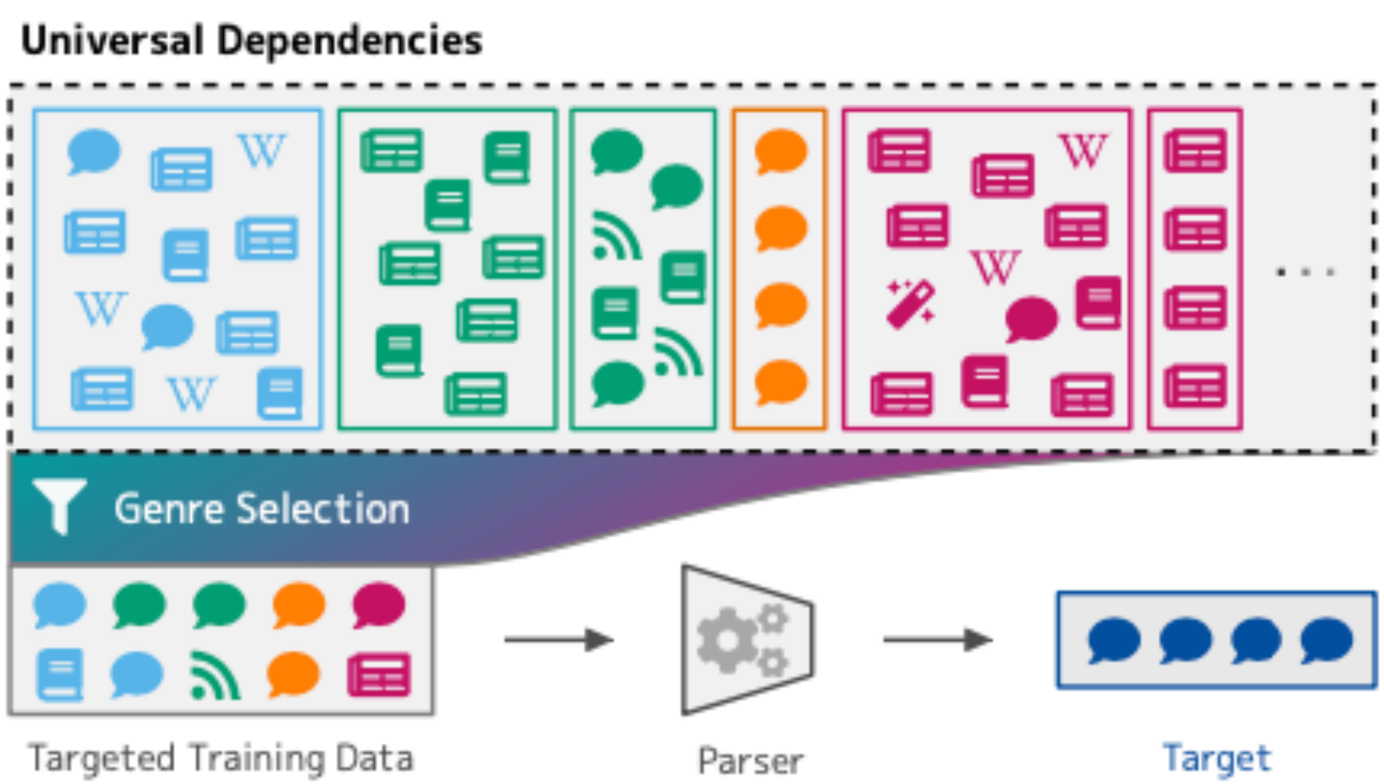
Data Selection for Low-resource Parsing

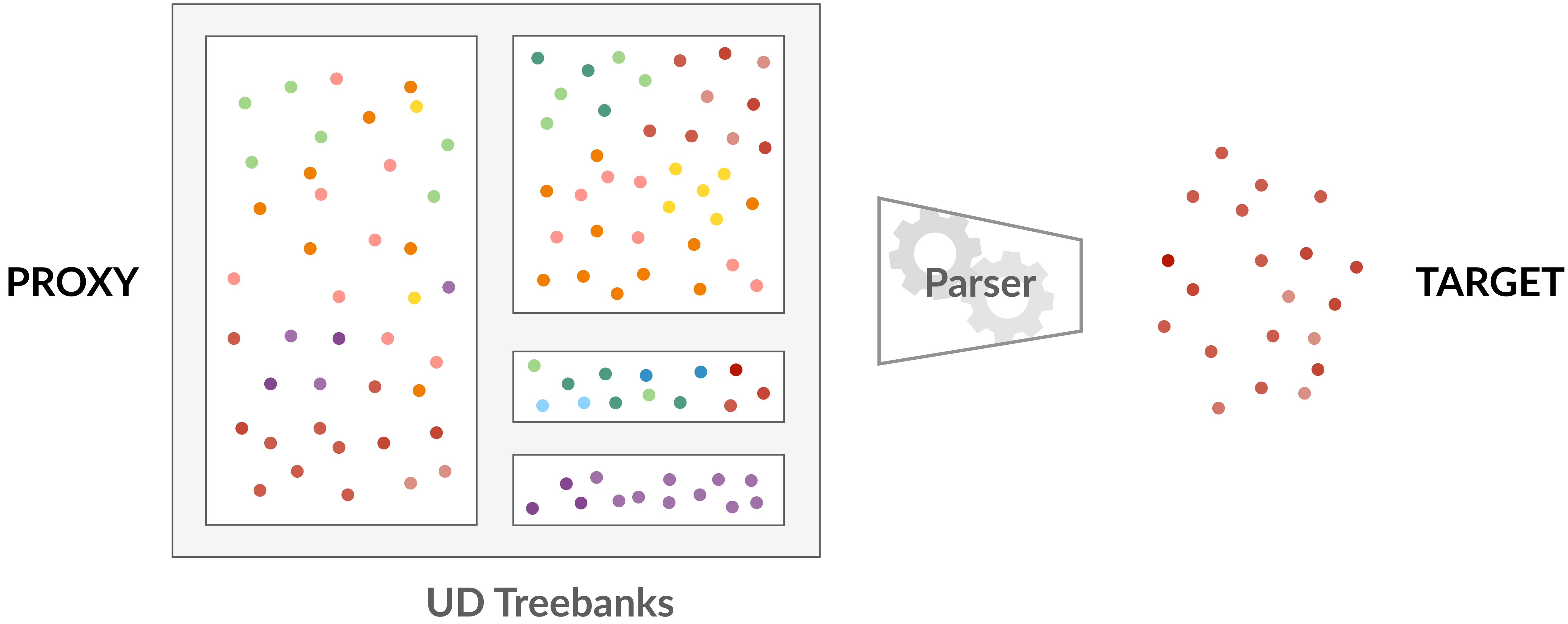
- **Problem & Motivation:**
  - A single parser trained on 100+ languages is suboptimal (training time, accuracy); also: for a practitioner it is difficult to choose appropriate training material.
  - Given UD, can we find better targeted training data?

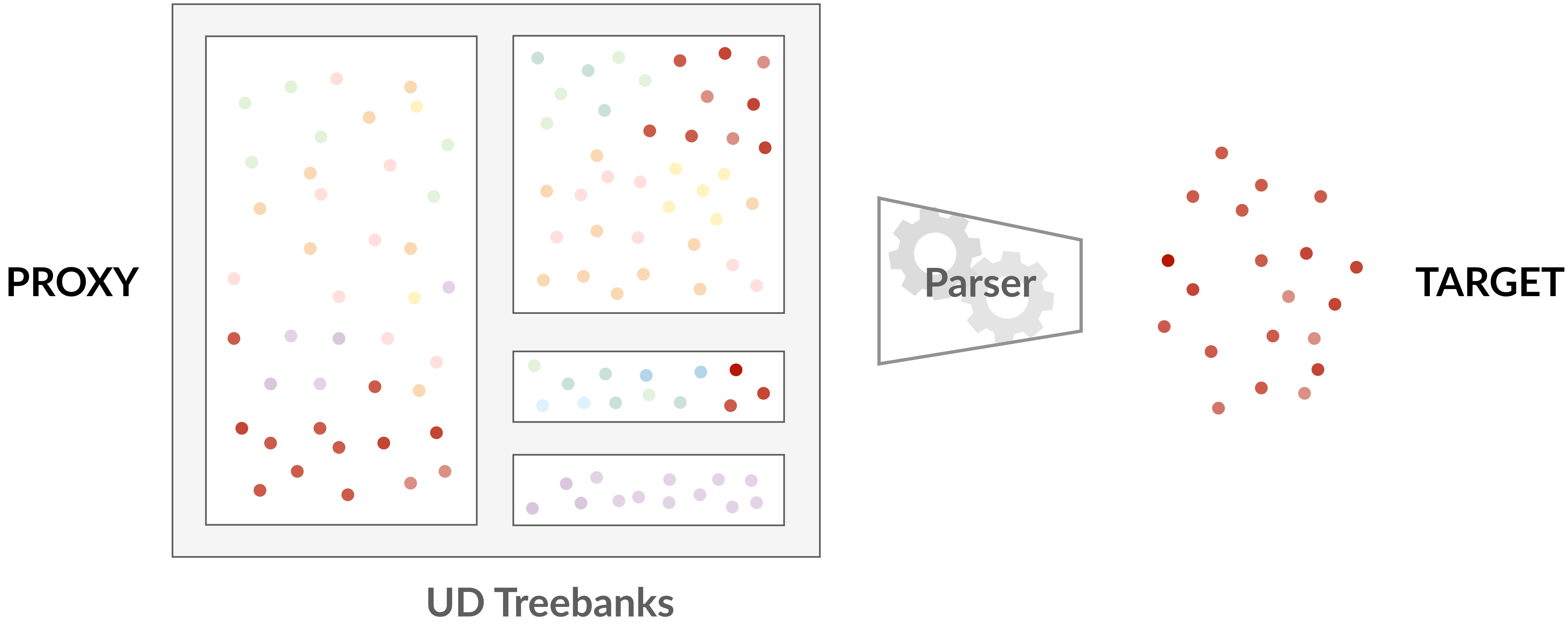


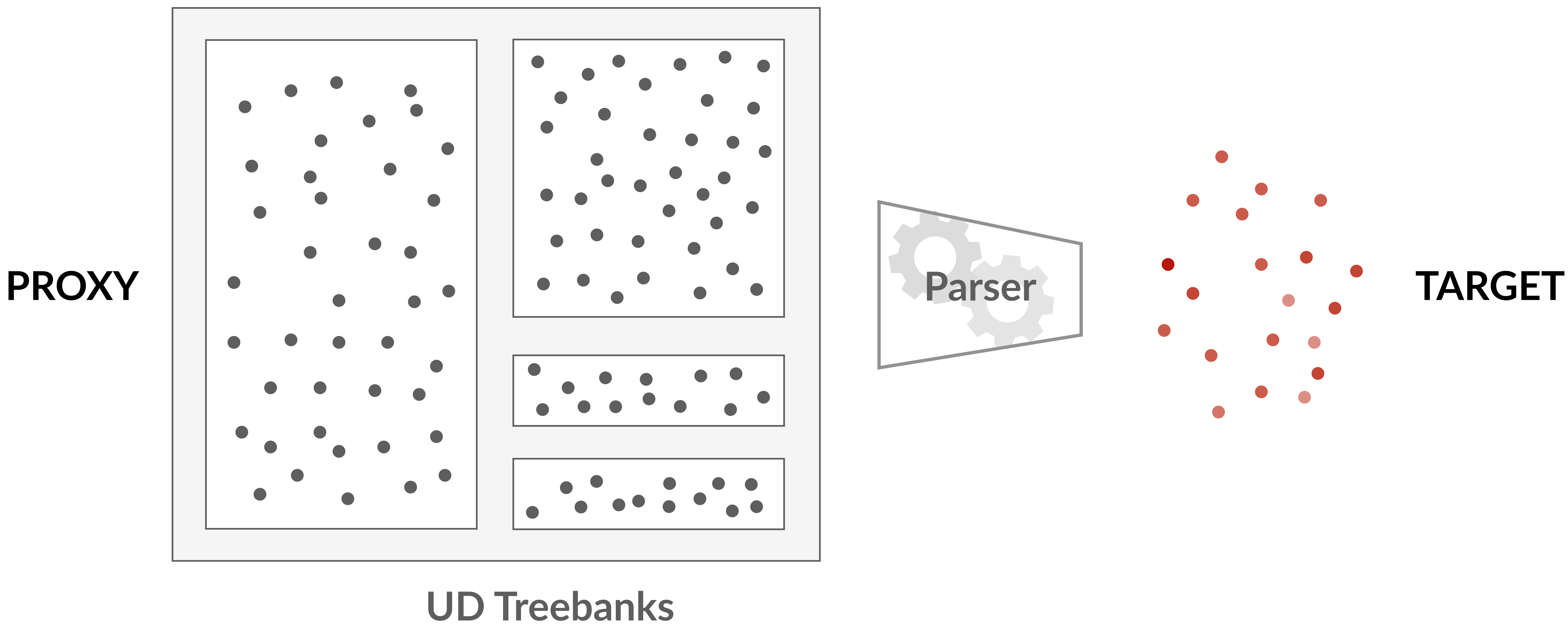
Key Idea: Genre as Fortuitous treebank-level meta-data

- **Research Questions:**
  - RQ1: To what extent does **genre** aid better proxy target data?
  - RQ2: Is genre **inherently** captured in multilingual LMs?




PROXY
Parser
TARGET
UD Treebanks


PROXY
UD Treebanks
Parser
TARGET


PROXY
UD Treebanks
Parser
TARGET

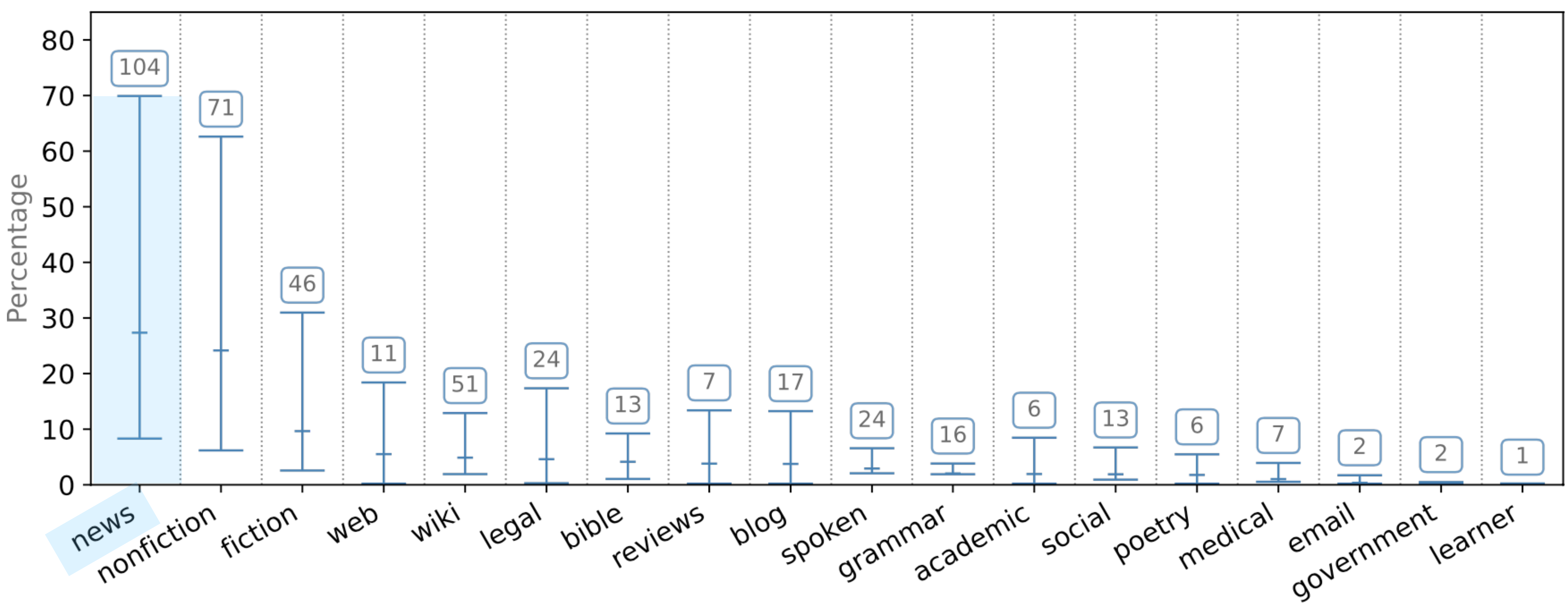
Genre as Weak Supervision

Domain $\boxed{\textbf{Genre}}$ Register

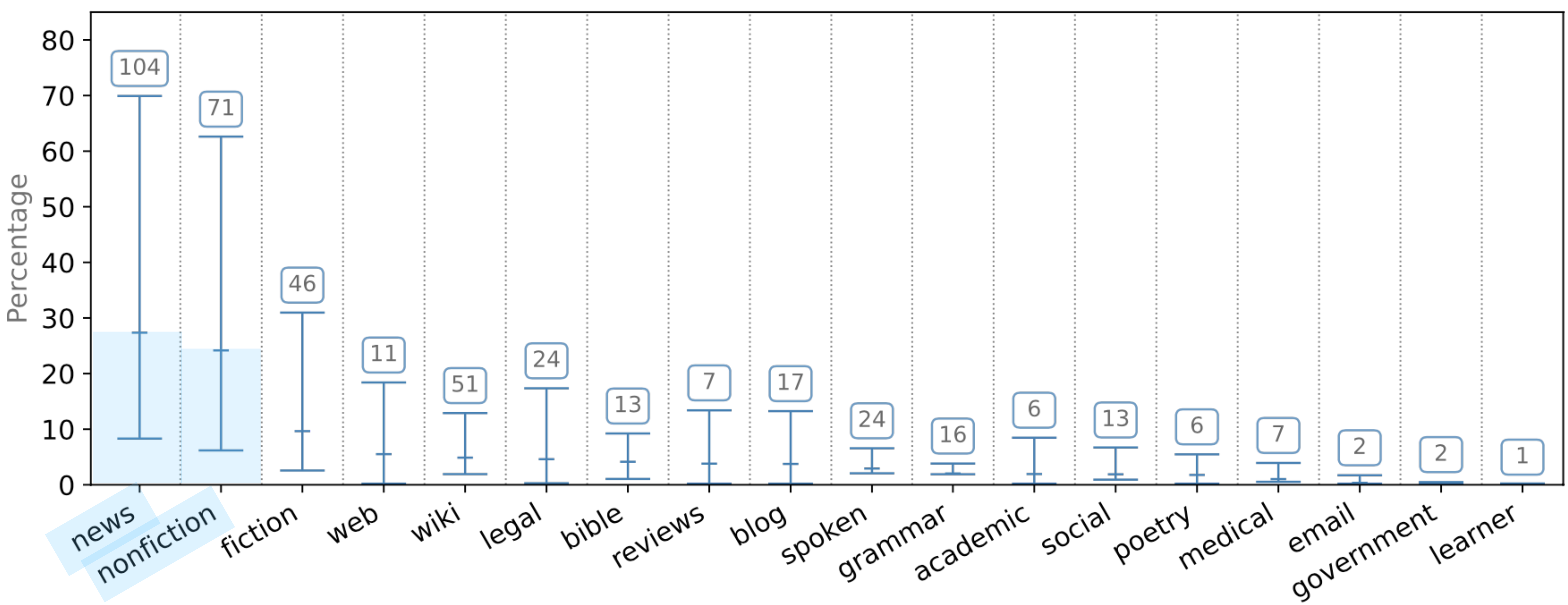
Kessler et al. (1997); Lee (2001); Webber (2009); Plank (2011)

18 community-provided categories in UD

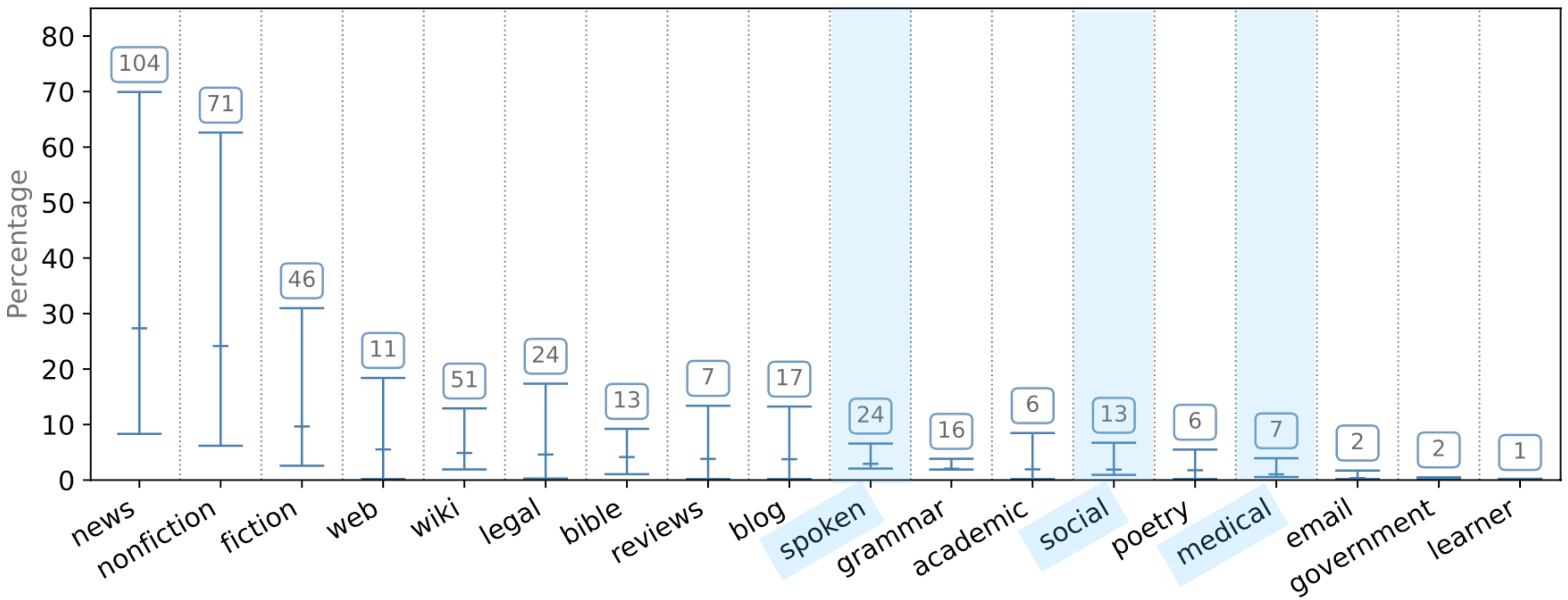
Genre Distribution in UD



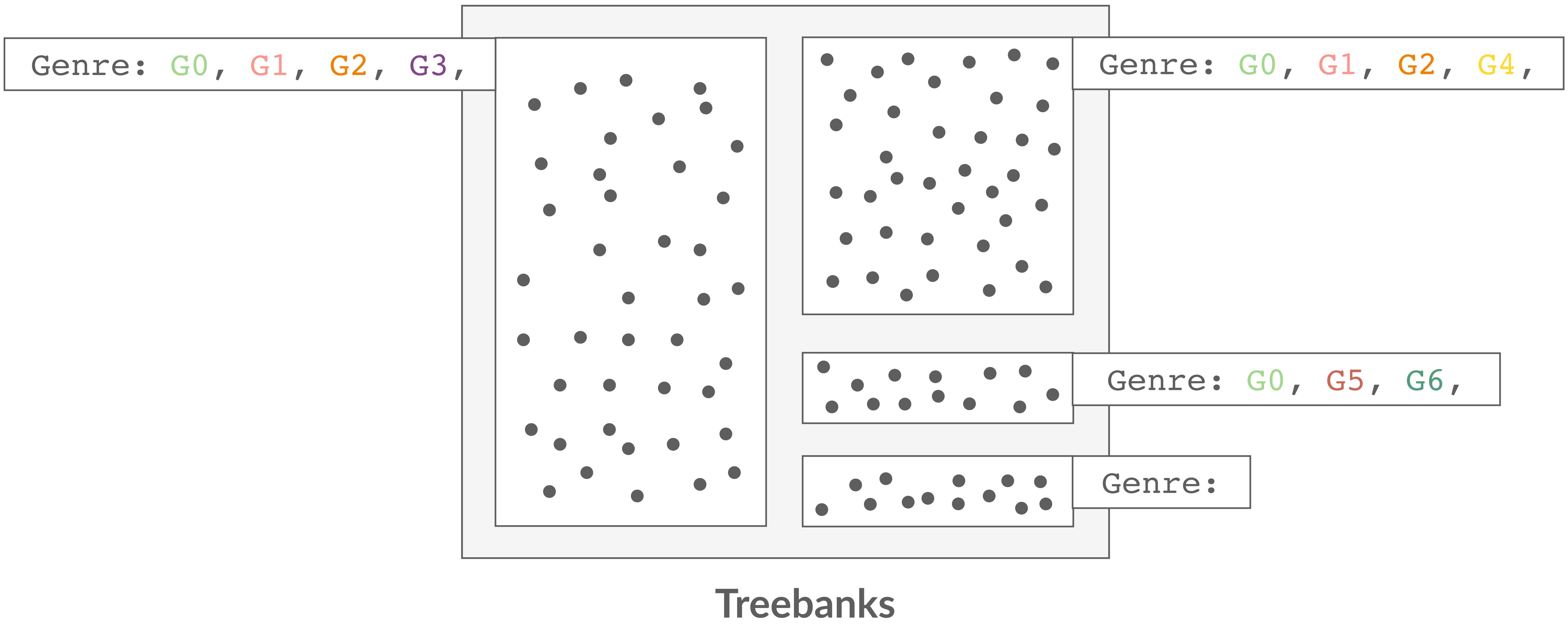
Genre Distribution in UD



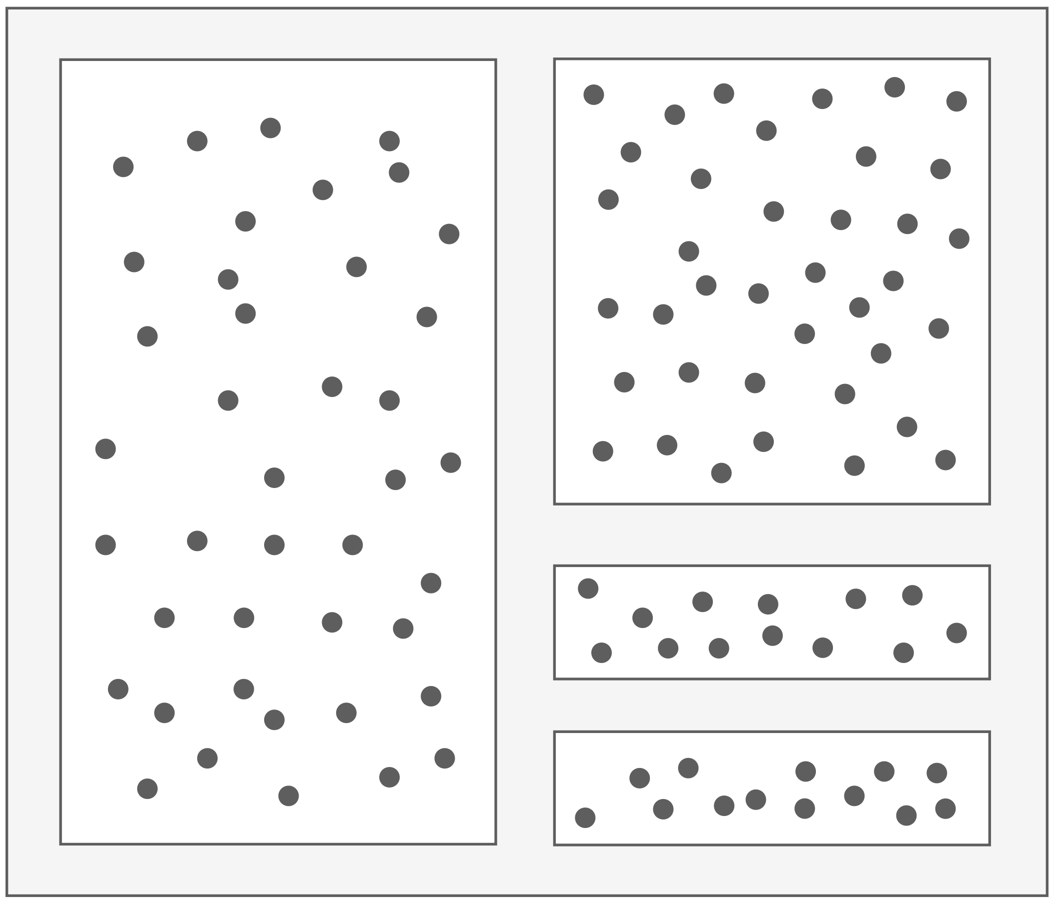
Genre Distribution in UD



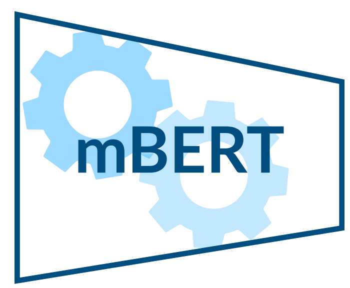
Targeted Data Selection


Genre: G0, G1, G2, G3,
Genre: G0, G1, G2, G4,
Genre: G0, G5, G6,
Genre:
Treebanks

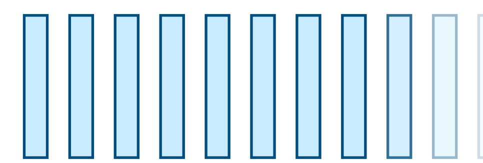
	MODEL	GENRES	LANGS
This Work	`mBERT`	18	104
Aharoni & Goldberg (2020)	`BERT`	5	1



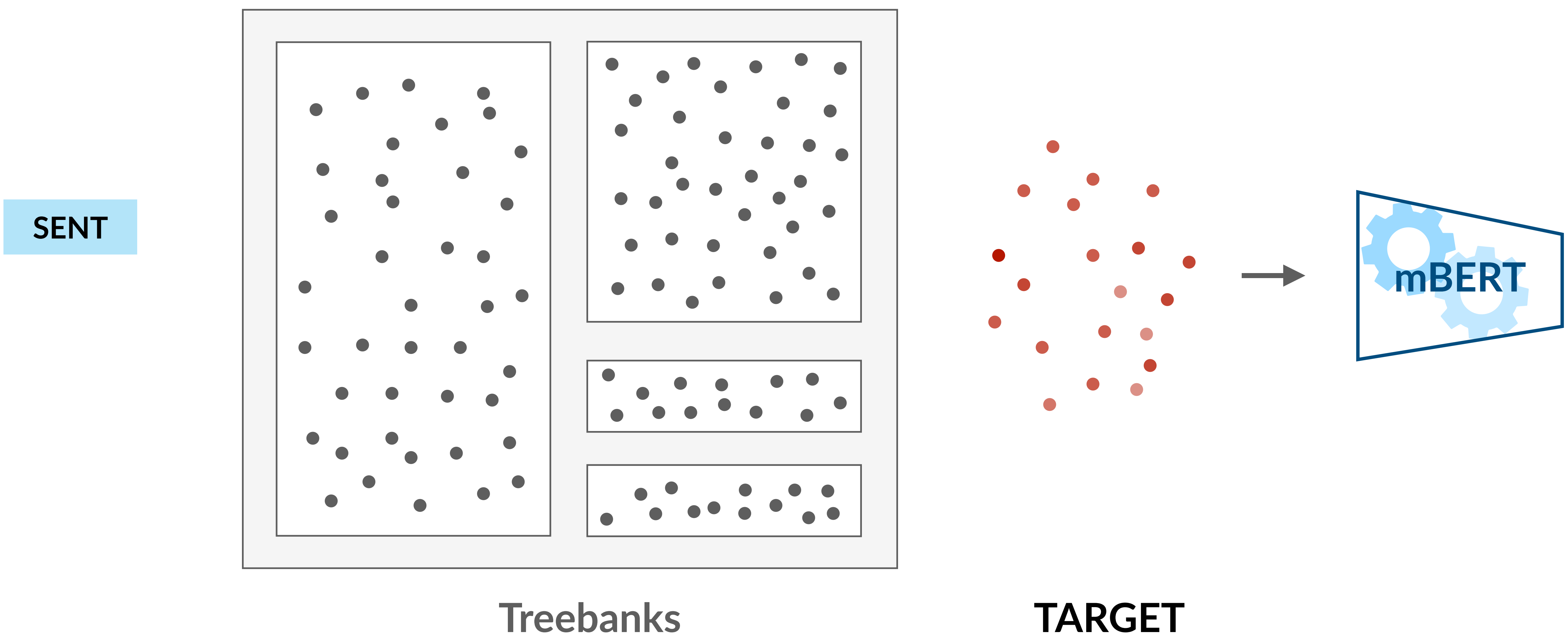
Treebanks



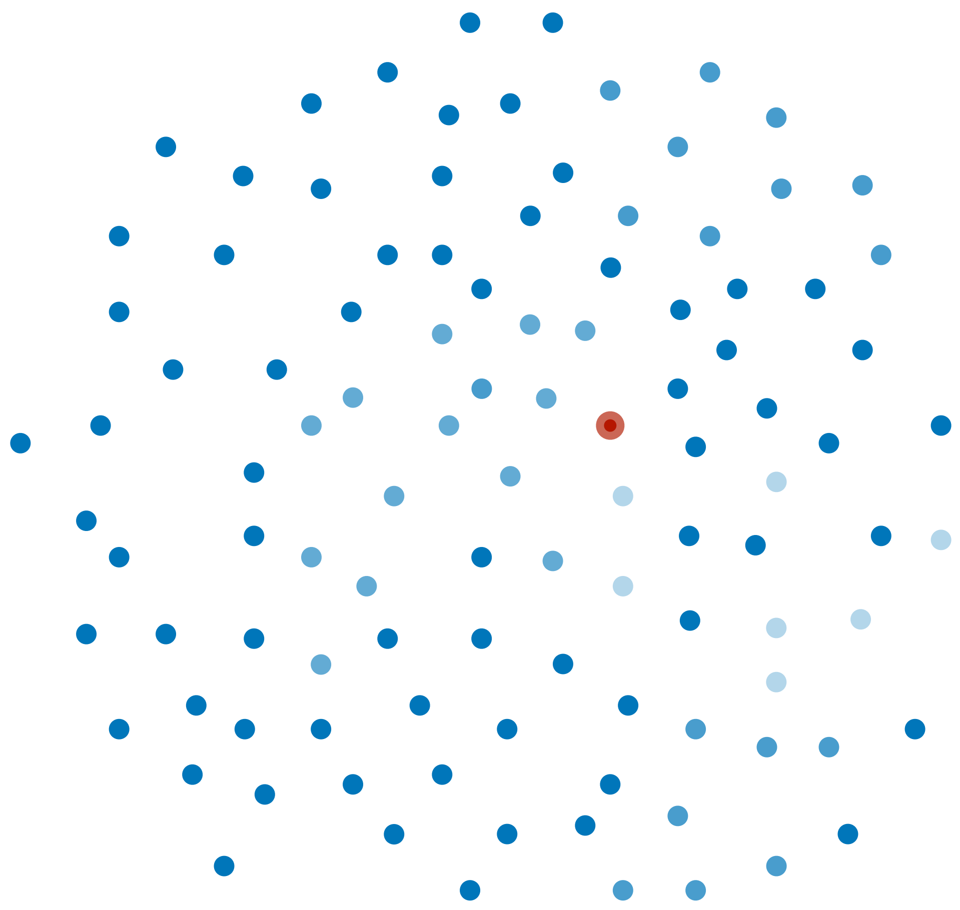
Devlin et al. (2019)



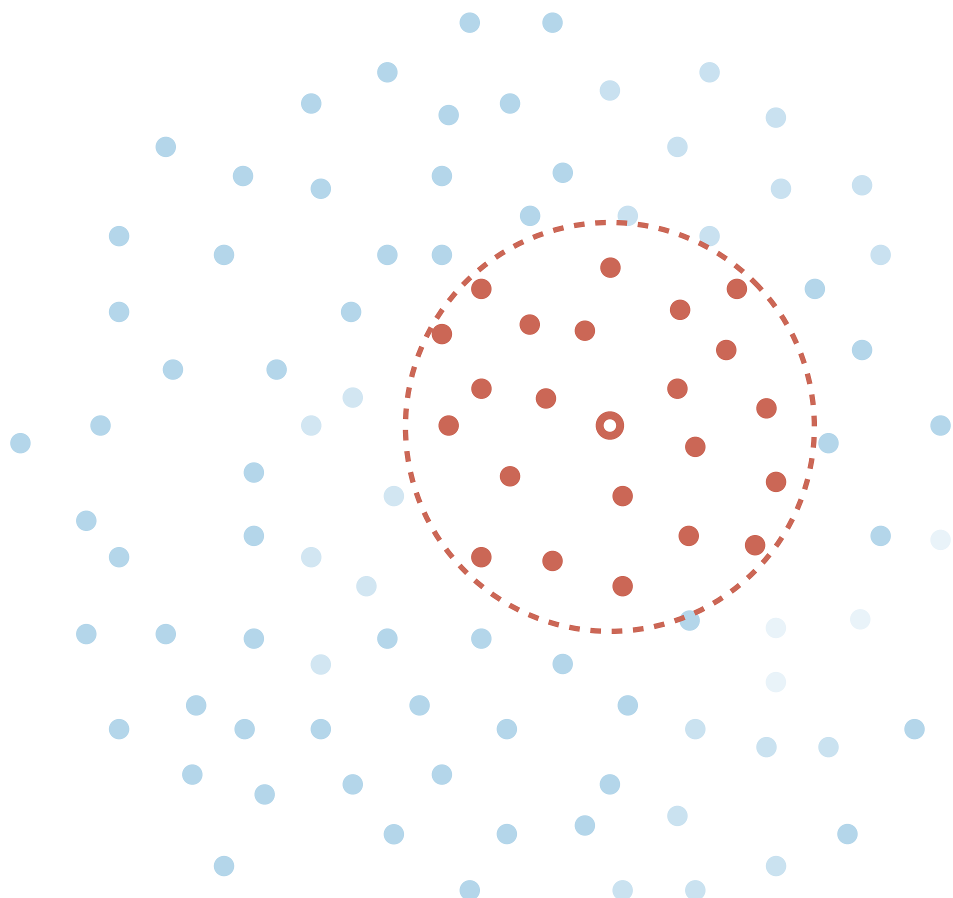
SENT


SENT
Treebanks
TARGET
mBERT

SENT



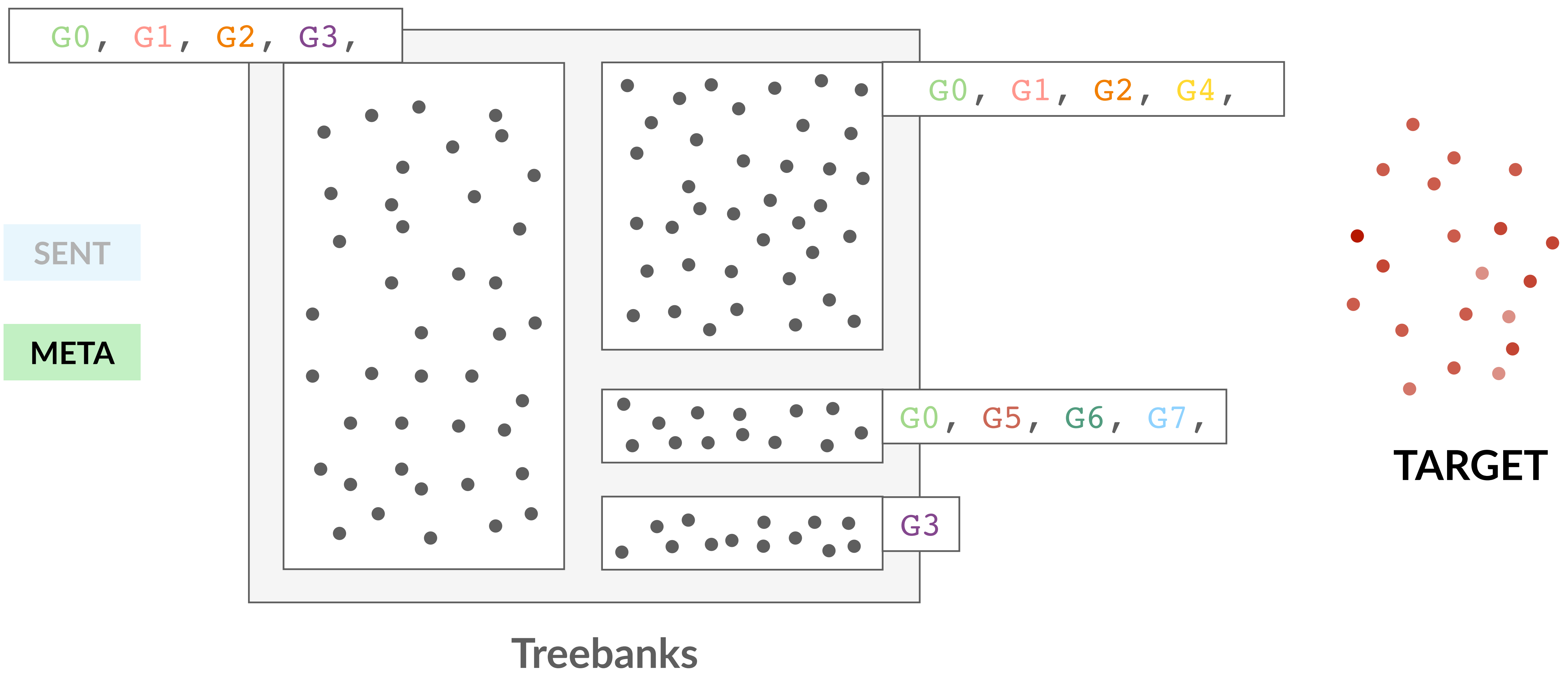
SENT

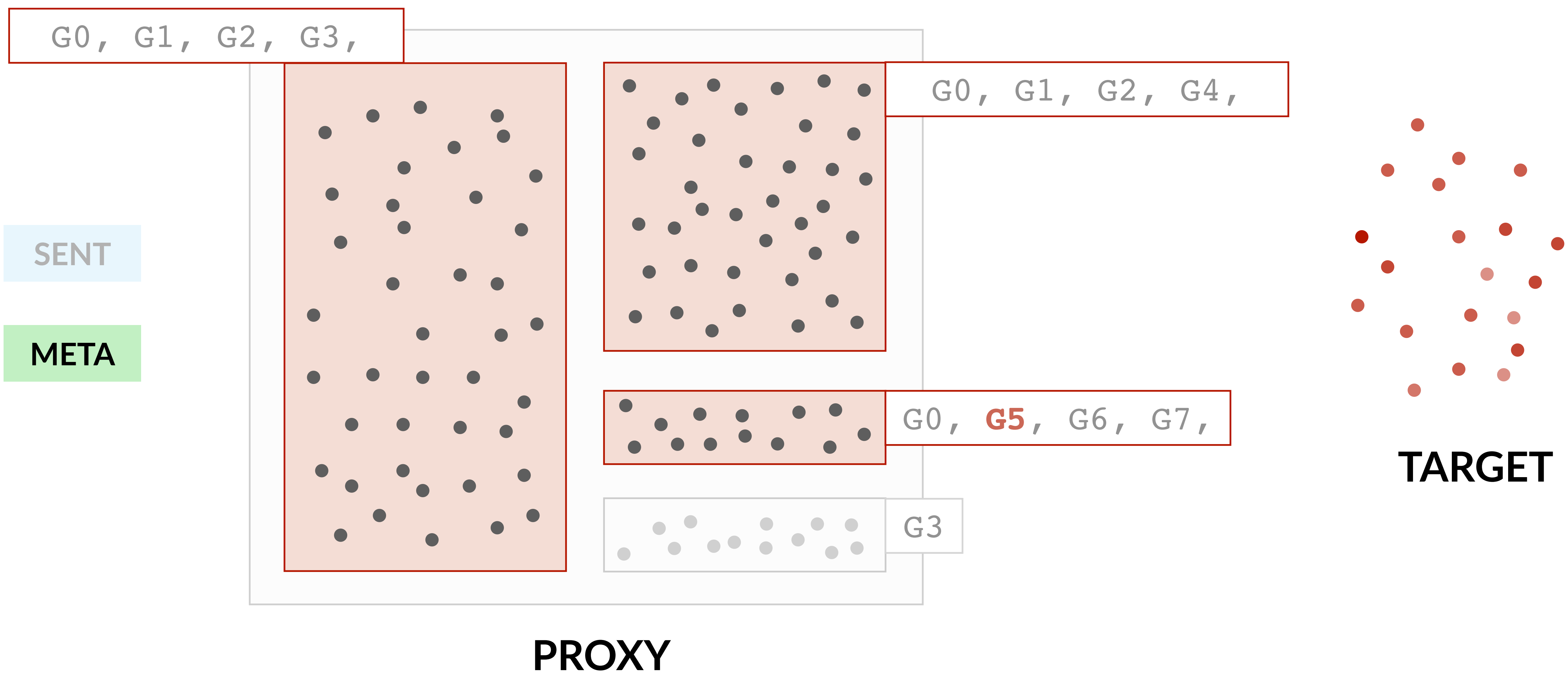


PROXY

SENT

META


G0, G1, G2, G3,
SENT
META
G0, G1, G2, G4,
G0, G5, G6, G7,
G3
Treebanks
TARGET

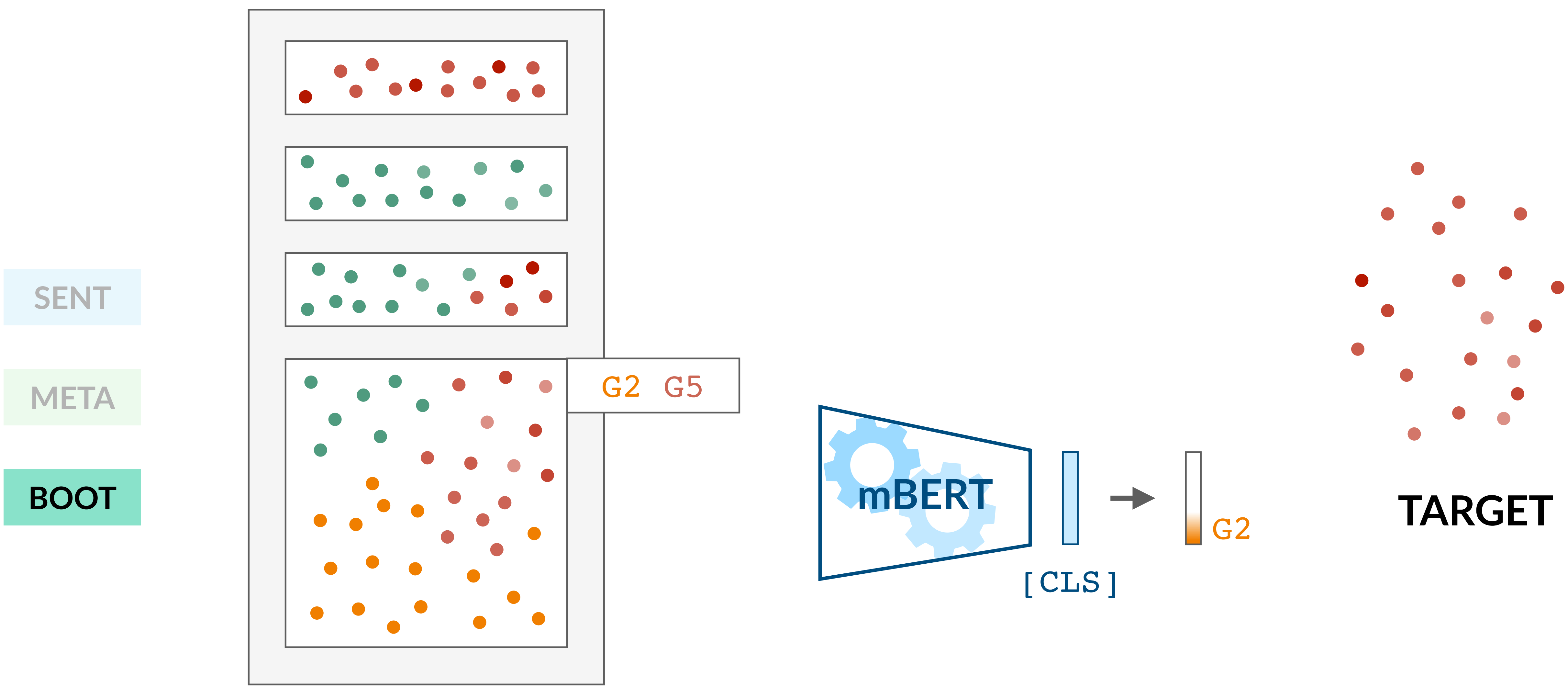


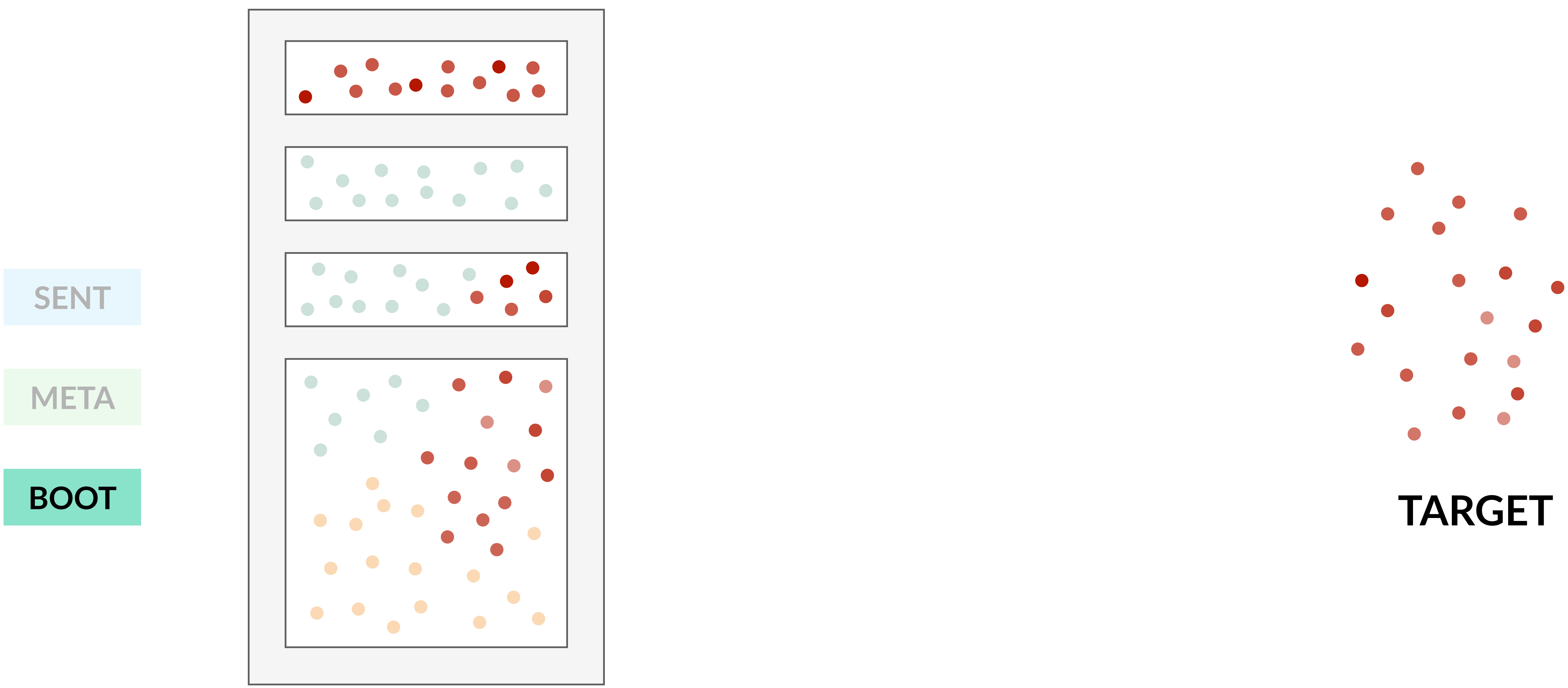
SENT
META
G0, G1, G2, G3,
G0, G1, G2, G4,
G0, G5, G6, G7,
G3
PROXY
TARGET

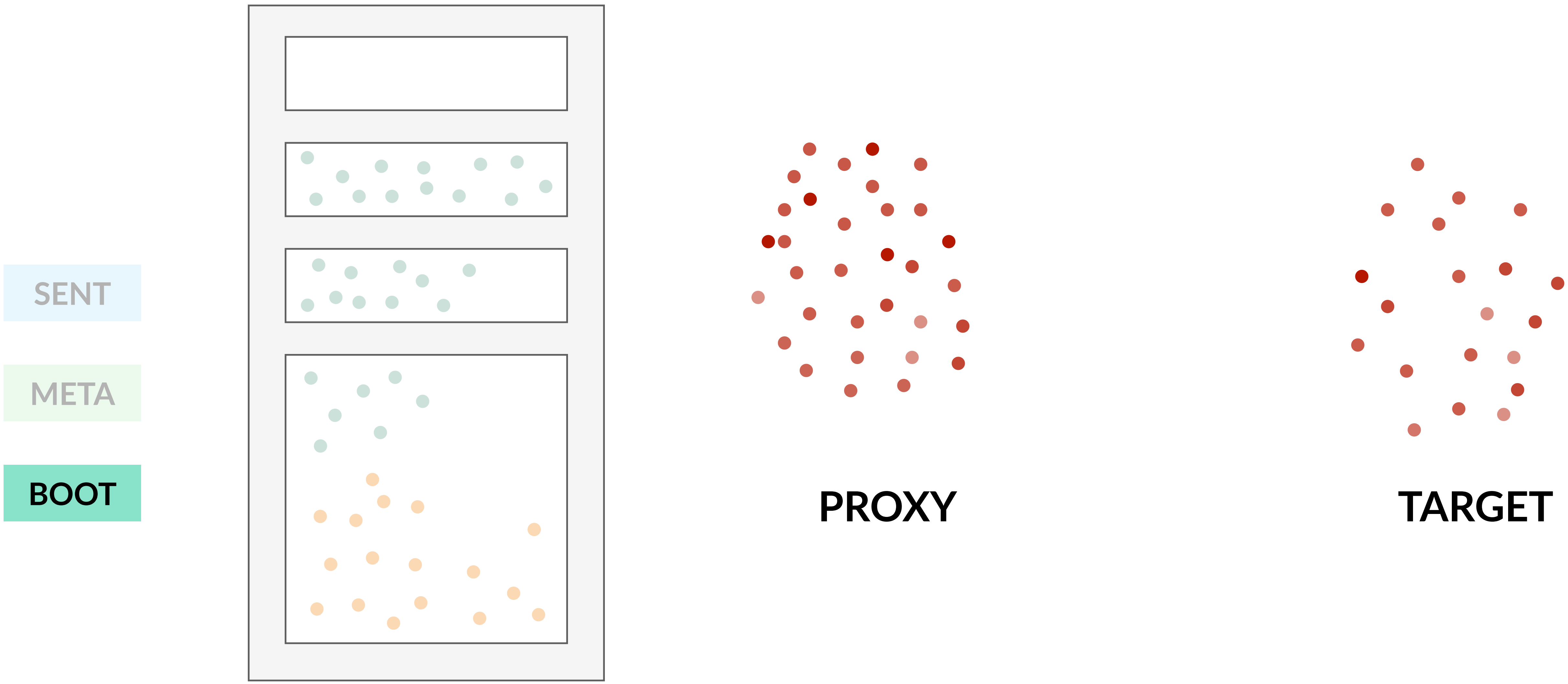
SENT

META

BOOT


SENT
META
BOOT
G2 G5
mBERT
[CLS]
G2
TARGET


SENT
META
BOOT
TARGET


SENT
META
BOOT
PROXY
TARGET

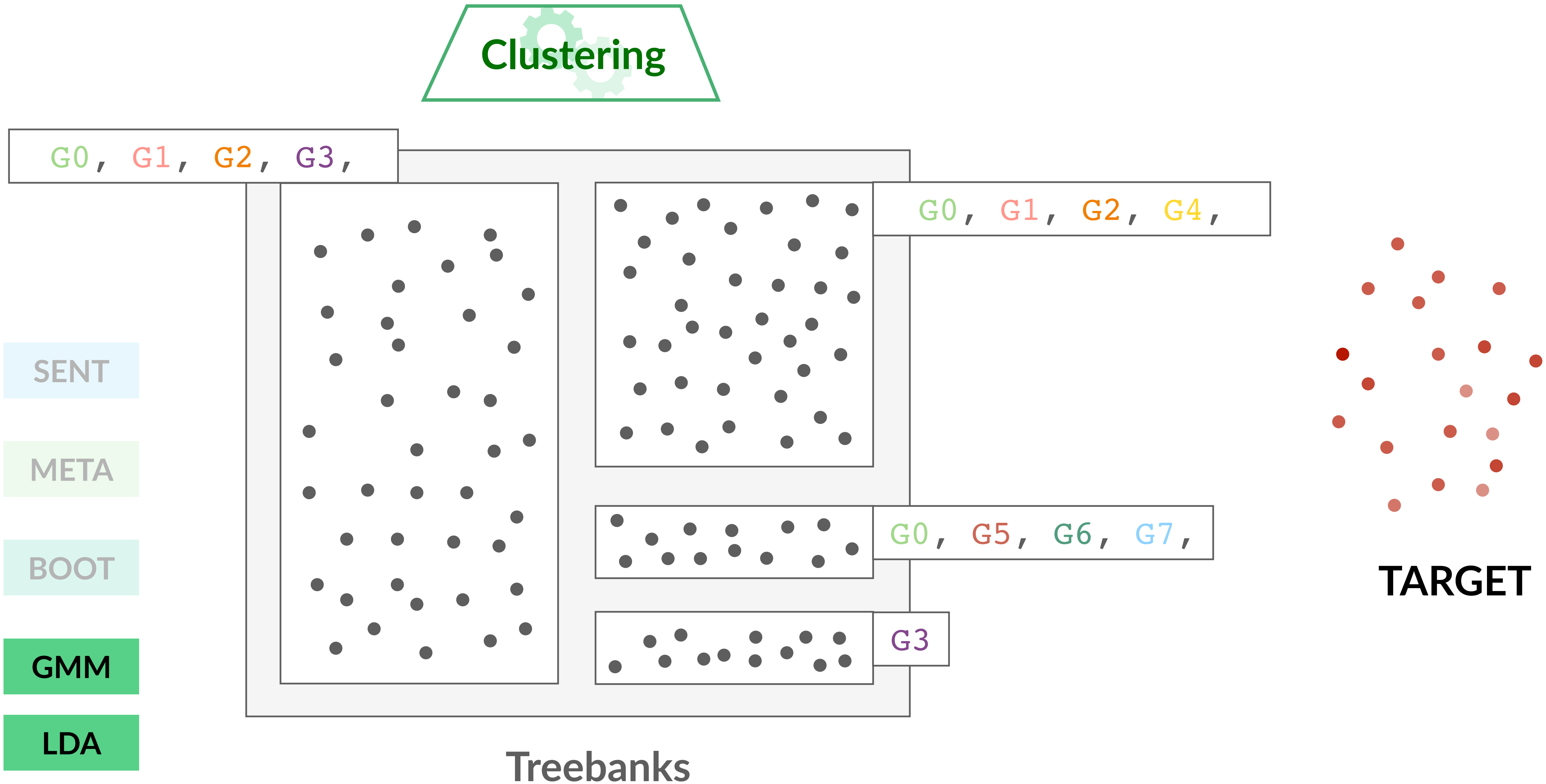
SENT

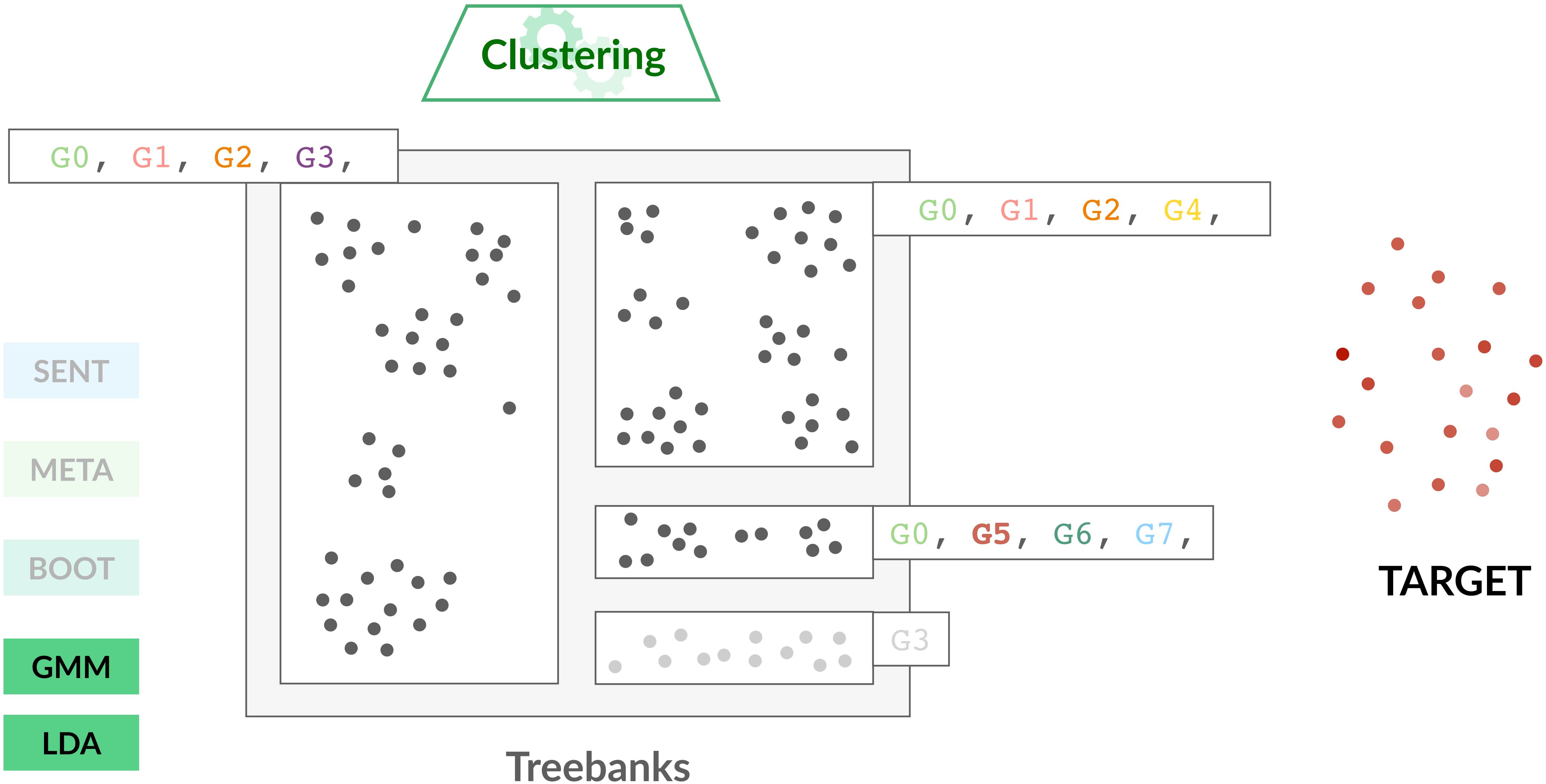
META

BOOT

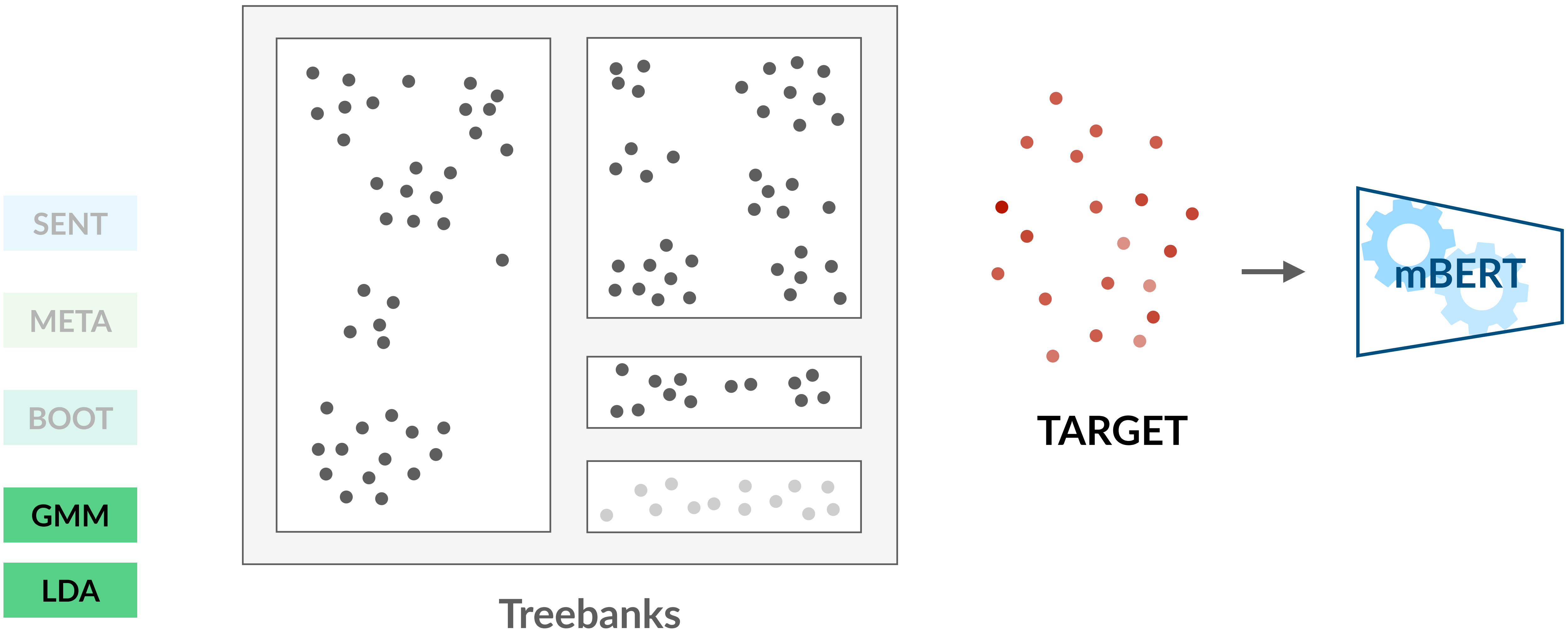
GMM

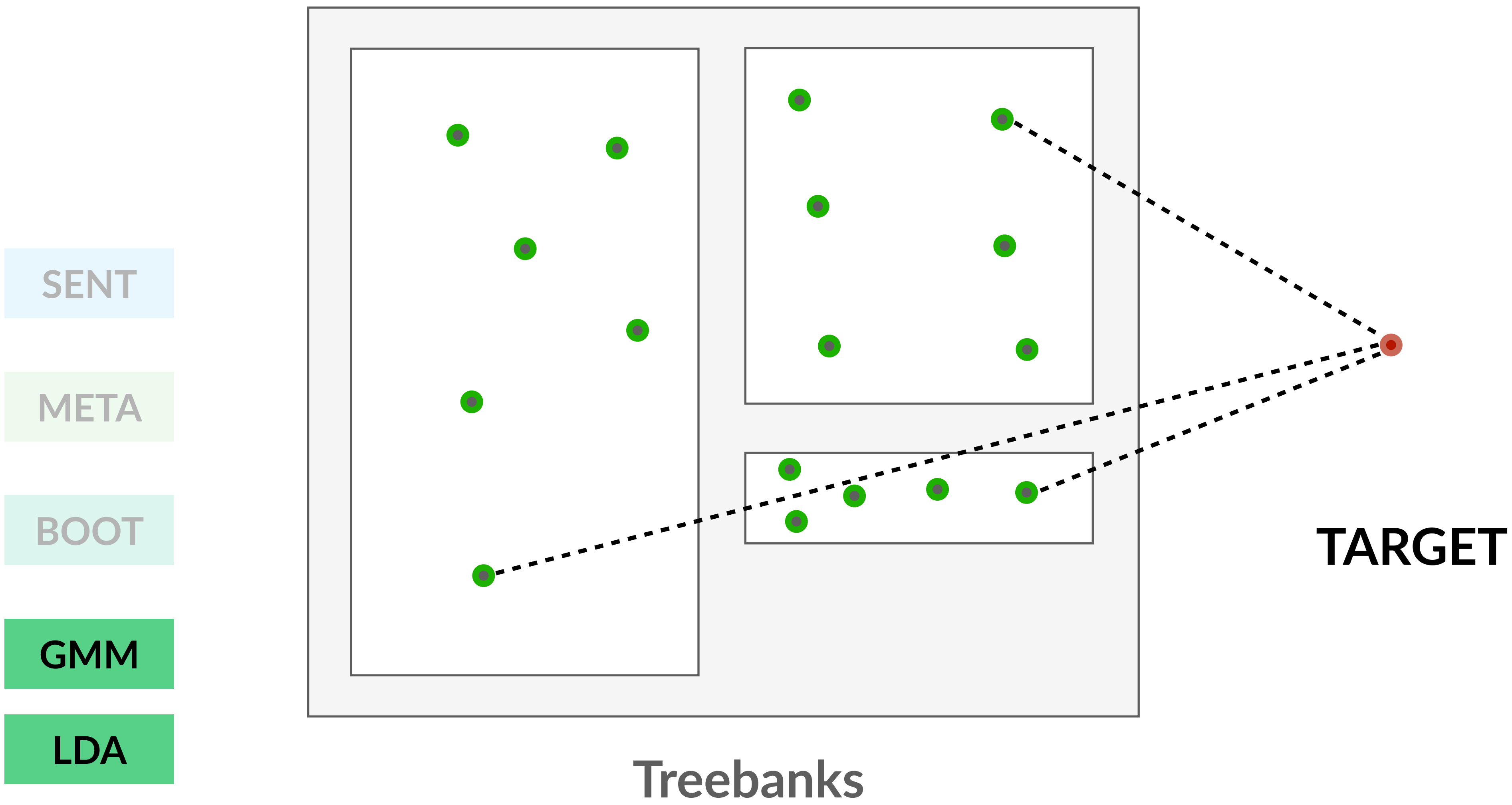
LDA

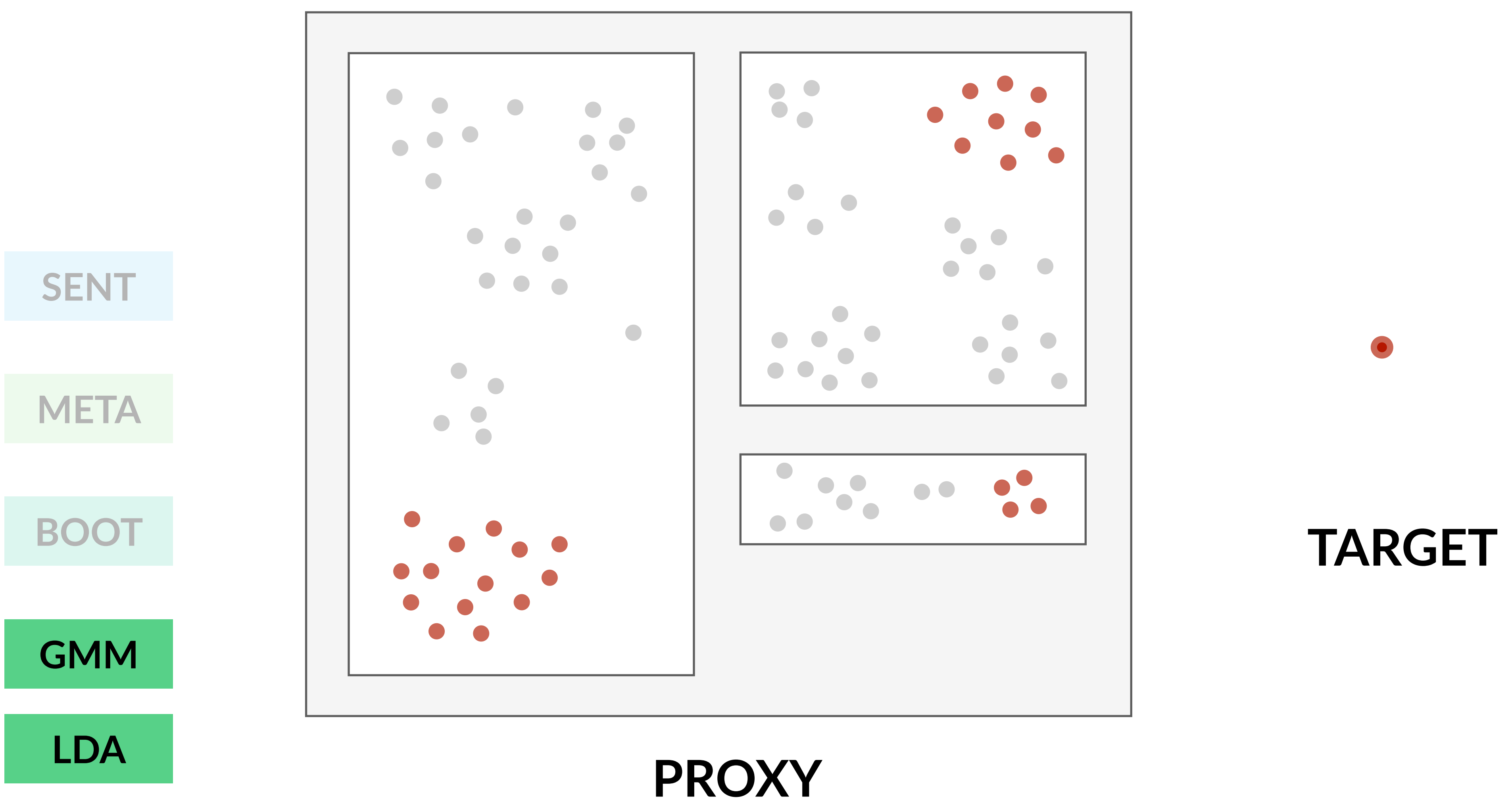

Clustering
G0, G1, G2, G3,
G0, G1, G2, G4,
G0, G5, G6, G7,
G3
SENT
META
BOOT
GMM
LDA
Treebanks
TARGET



Clustering
G0, G1, G2, G3,
G0, G1, G2, G4,
G0, G5, G6, G7,
G3
SENT
META
BOOT
GMM
LDA
Treebanks
TARGET


SENT
META
BOOT
GMM
LDA
Treebanks
TARGET
mBERT


SENT
META
BOOT
GMM
LDA
TARGET
Treebanks


SENT
META
BOOT
GMM
LDA
PROXY
TARGET

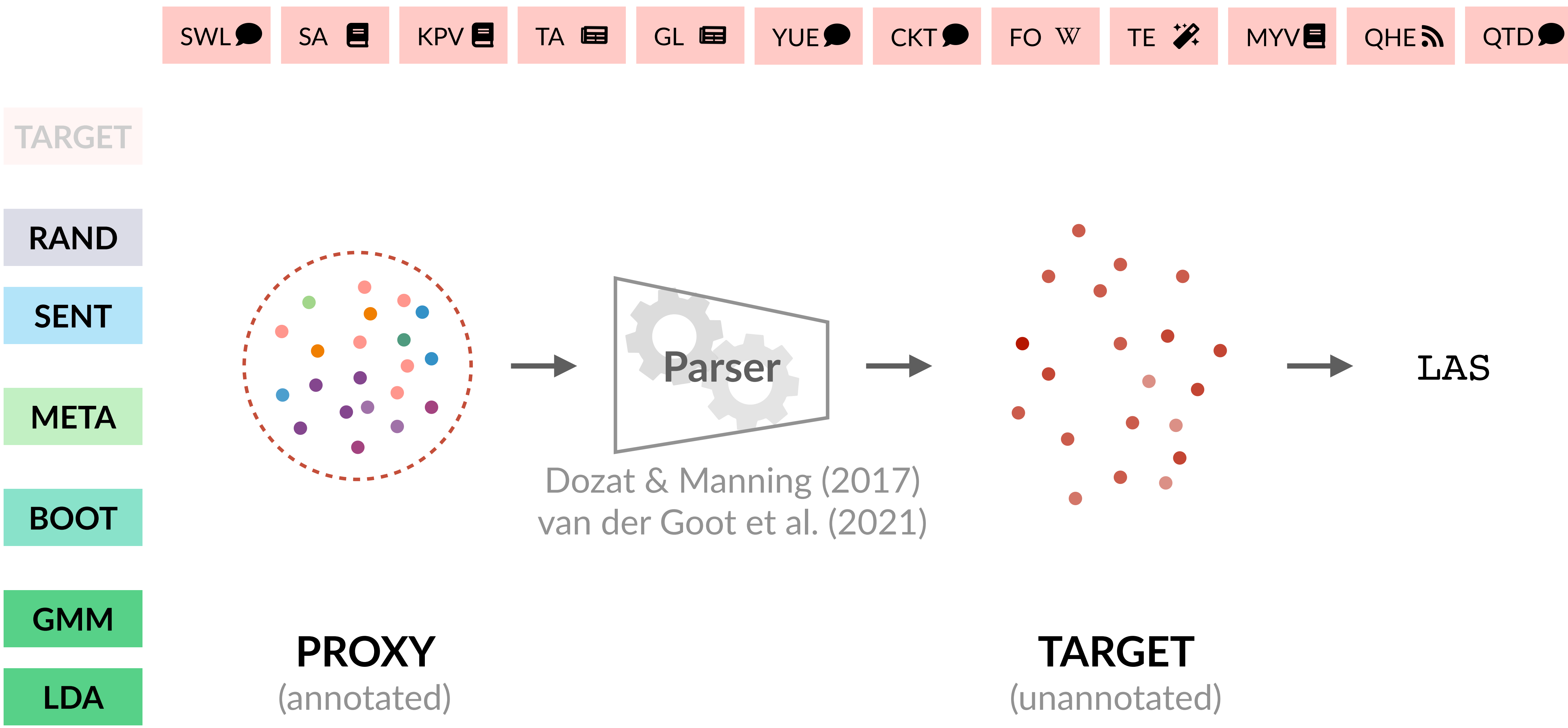
Experiments

Target		Authors	Language	#Sentences	mBERT	Genre
SWL	SSLC	Östling et al. (2017)	Swedish Sign Language	203	✗	spoken
SA	UFAL	Dwivedi and Easha (2017)	Sanskrit	230	✗	fiction
KPV	Lattice	Partanen et al. (2018)	Komi Zyrian	435	✗	fiction
TA	TTB	Ramasamy & Žabokrtský (2012)	Tamil	600	✓	news
GL	TreeGal	Garcia (2016)	Galician	1,000	✓	news
YUE	HK	Wong et al. (2017)	Cantonese	1,004	✗	spoken
CKT	HSE	Tyers and Mishchenkova (2020)	Chukchi	1,004	✗	spoken
FO	OFT	Tyers et al. (2018)	Faroese	1,208	✗	wiki
TE	MTG	Rama and Vajjala (2017)	Telugu	1,328	✓	grammar
MYV	JR	Rueter and Tyers (2018)	Erzya	1,690	✗	fiction
QHE	HIENCS	Bhat et al. (2018)	Hindi-English	1,800	~	social
QTD	SAGT	Çetinoğlu and Çöltekin (2019)	Turkish-German	1,891	~	spoken


SWL
SA
KPV
TA
GL
YUE
CKT
FO
TE
MYV
QHE
QTD
SENT
META
BOOT
GMM
LDA

	SWL	SA	KPV	TA	GL	YUE	CKT	FO	TE	MYV	QHE	QTD
TARGET	✓	~	~	✓	✓	✗	✗	~	✓	✗	✓	✓
SENT												
META												
BOOT												
GMM												
LDA												


SWL
SA
KPV
TA
GL
YUE
CKT
FO
TE
MYV
QHE
QTD
TARGET
RAND
SENT
META
BOOT
GMM
LDA

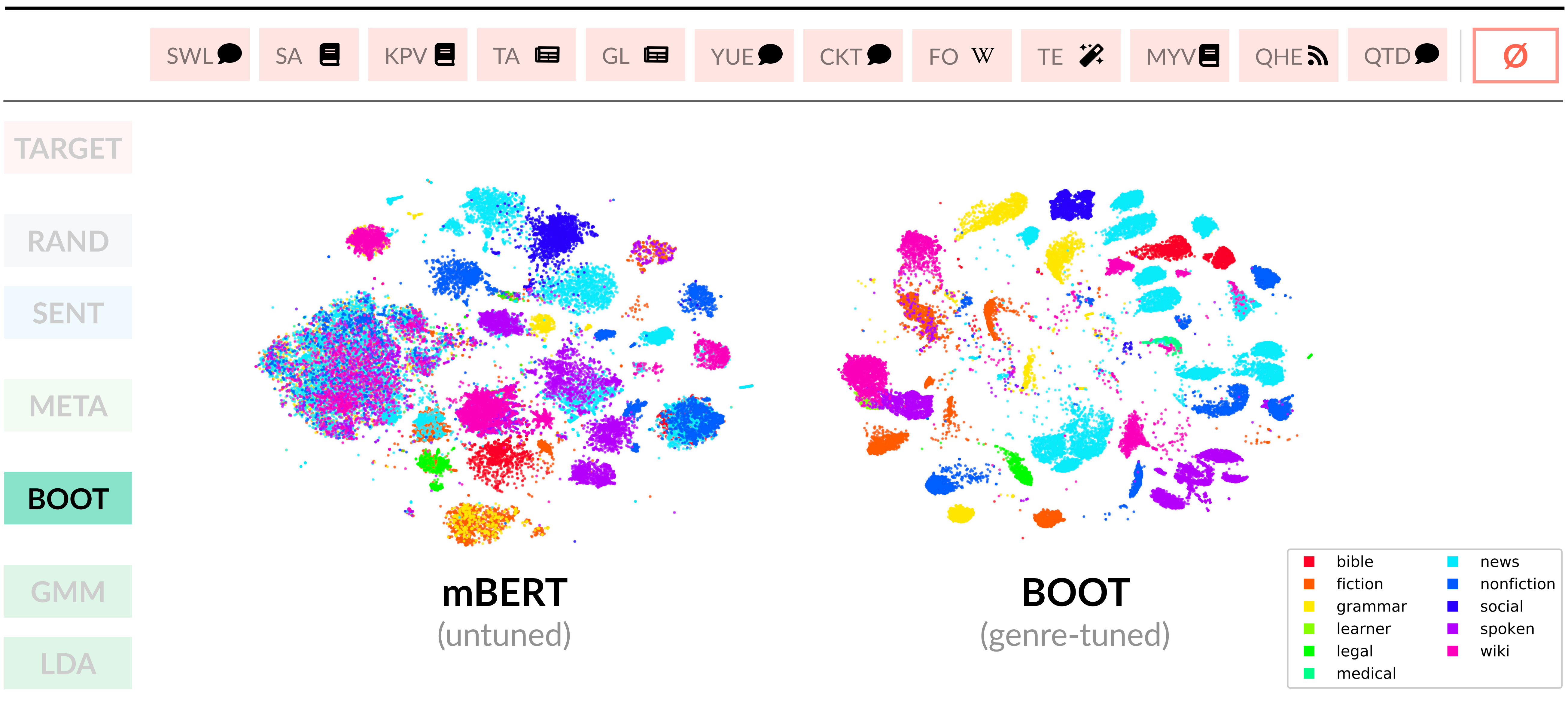

SWL
SA
KPV
TA
GL
YUE
CKT
FO
TE
MYV
QHE
QTD
TARGET
RAND
SENT
META
BOOT
GMM
LDA
Parser
Dozat & Manning (2017)
van der Goot et al. (2021)
LAS
PROXY
(annotated)
TARGET
(unannotated)

	SWL	SA	KPV	TA	GL	YUE	CKT	FO	TE	MYV	QHE	QTD	Ø
TARGET	28.0	15.7	13.4	64.1	80.9	—	—	49.6	83.6	—	62.7	55.0	50.3
RAND													
SENT													
META													
BOOT													
GMM													
LDA													

	SWB	SA	KPV	TA	GE	YUB	CKT	FO	TE	MYV	QHE	QTD	Ø
TARGET	28.0	15.7	13.4	64.1	80.9	—	—	49.6	83.6	—	62.7	55.0	50.3
RAND													
SENT													
META	6.5	24.3	10.2	50.4	76.6	31.2	11.6	61.2	64.9	20.4	9.42	42.6	34.1
BOOT													
GMM													
LDA													

	SWL	SA	KPV	TA	GL	YUE	CKT	FO	TE	MYV	QHE	QTD	Ø
TARGET	28.0	15.7	13.4	64.1	80.9	—	—	49.6	83.6	—	62.7	55.0	50.3
RAND	3.7	24.8	10.9	50.7	77.7	33.3	15.5	61.9	67.7	20.0	27.0	44.6	36.5
SENT	3.6	23.7	13.7	47.9	77.6	35.8	16.4	62.5	68.1	22.9	26.5	42.8	36.8
META	6.5	24.3	10.2	50.4	76.6	31.2	11.6	61.2	64.9	20.4	9.42	42.6	34.1
BOOT													
GMM													
LDA													

	SWL	SA	KPV	TA	GL	YUE	CKT	FO	TE	MYV	QHE	QTD	Ø
TARGET	28.0	15.7	13.4	64.1	80.9	—	—	49.6	83.6	—	62.7	55.0	50.3
RAND	3.7	**24.8**	10.9	50.7	77.7	33.3	15.5	61.9	67.7	20.0	**27.0**	44.6	36.5
SENT	3.6	23.7	13.7	47.9	77.6	35.8	16.4	62.5	68.1	**22.9**	26.5	42.8	36.8
META	6.5	24.3	10.2	50.4	76.6	31.2	11.6	61.2	64.9	20.4	9.42	42.6	34.1
BOOT	5.2	21.8	*21.1	49.4	76.7	*49.9	18.4	*66.3	65.6	19.5	14.8	43.8	37.7
GMM	4.9	22.9	*20.9	***51.5**	**77.8**	***49.9**	***19.8**	*68.3	67.9	20.2	15.1	**45.4**	**38.7**
LDA	**6.6**	23.7	***22.3**	49.2	77.0	*49.4	*19.1	***68.3**	***68.6**	20.5	15.1	44.7	**38.7**



SWL
SA
KPV
TA
GL
YUE
CKT
FO
TE
MYV
QHE
QTD
Ø
TARGET
RAND
SENT
META
BOOT
GMM
LDA
mBERT
(untuned)
BOOT
(genre-tuned)
bible
fiction
grammar
learner
legal
medical
news
nonfiction
social
spoken
wiki

Take-Aways

BOOT

GMM

LDA

RQ1: Genre is a valuable signal for parsing unseen, low-resource targets

RQ2: Genre is inherently captured in multilingual LMs and amplifying it helps to improve parsing performance

Roadmap

1. How useful is (fortuitous) meta-data for low-res parsing?
2. How impactful are segment embeddings for low-res NLP?
3. To what extent does auxiliary data help limited training data?

Frustratingly Easy Performance Improvements for Cross-lingual Transfer: A Tale on BERT and Segment Embeddings

Rob van der Goot,[♣] Max Müller-Eberstein,[♣] Barbara Plank[♣◇]
[♣]Computer Science Department, IT University of Copenhagen
[◇]Center for Information and Language Processing (CIS), LMU Munich, Germany
robv@itu.dk, mamy@itu.dk, bapl@itu.dk

Part 2

LREC, 2022

Segments: An understudied BERT detail?

➡ Question/Answer or Sentence follows (NSP)

On the Impact of Segment Embeddings

- We contribute an analysis of segment embeddings (for BERTology)
- **Research Questions:**
  - RQ1: To what extent does the choice of segment embedding (0,1) impact downstream performance?
  - RQ2: Are paired-sentence tasks more affected by segment IDs?

Segment Embeddings Variants

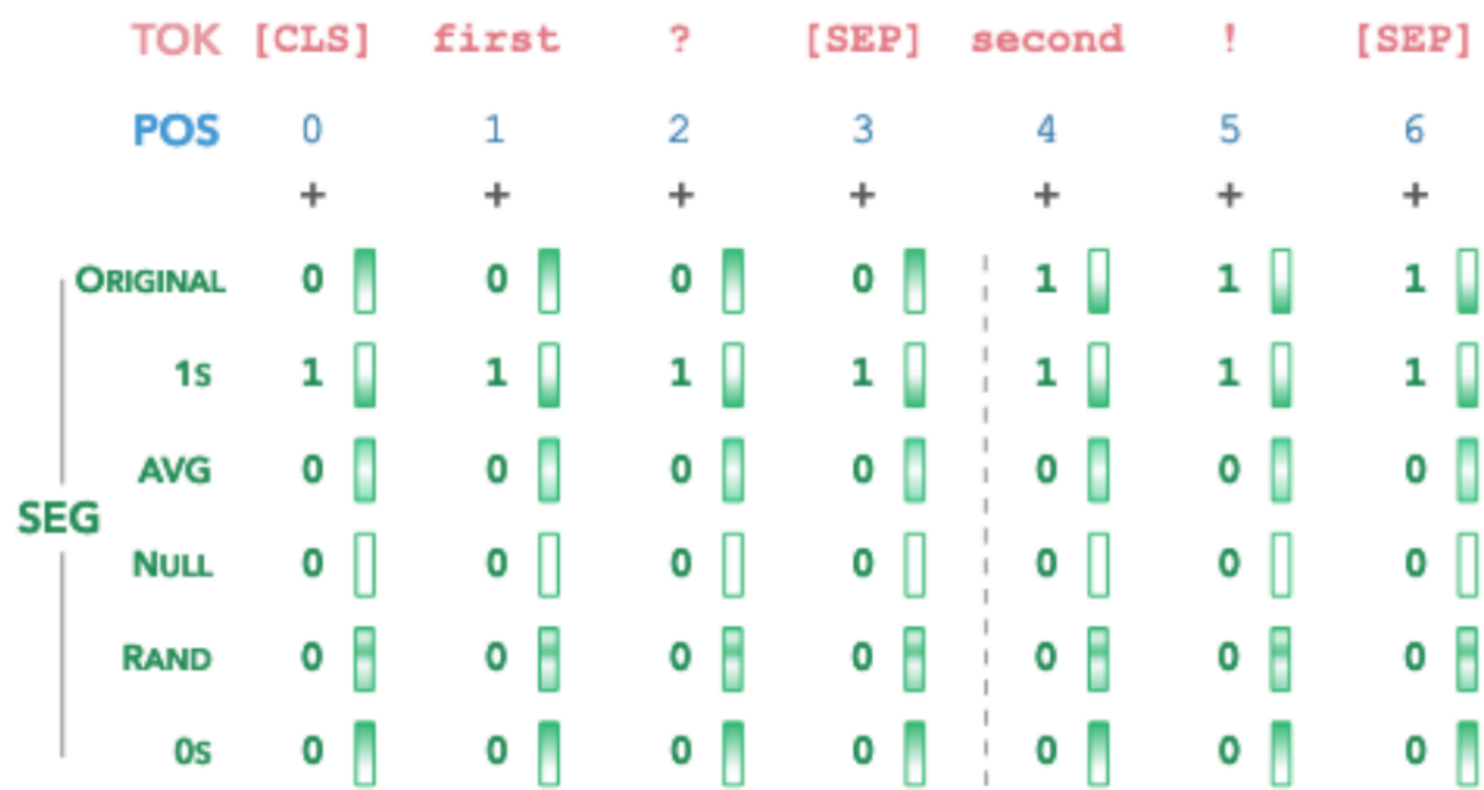


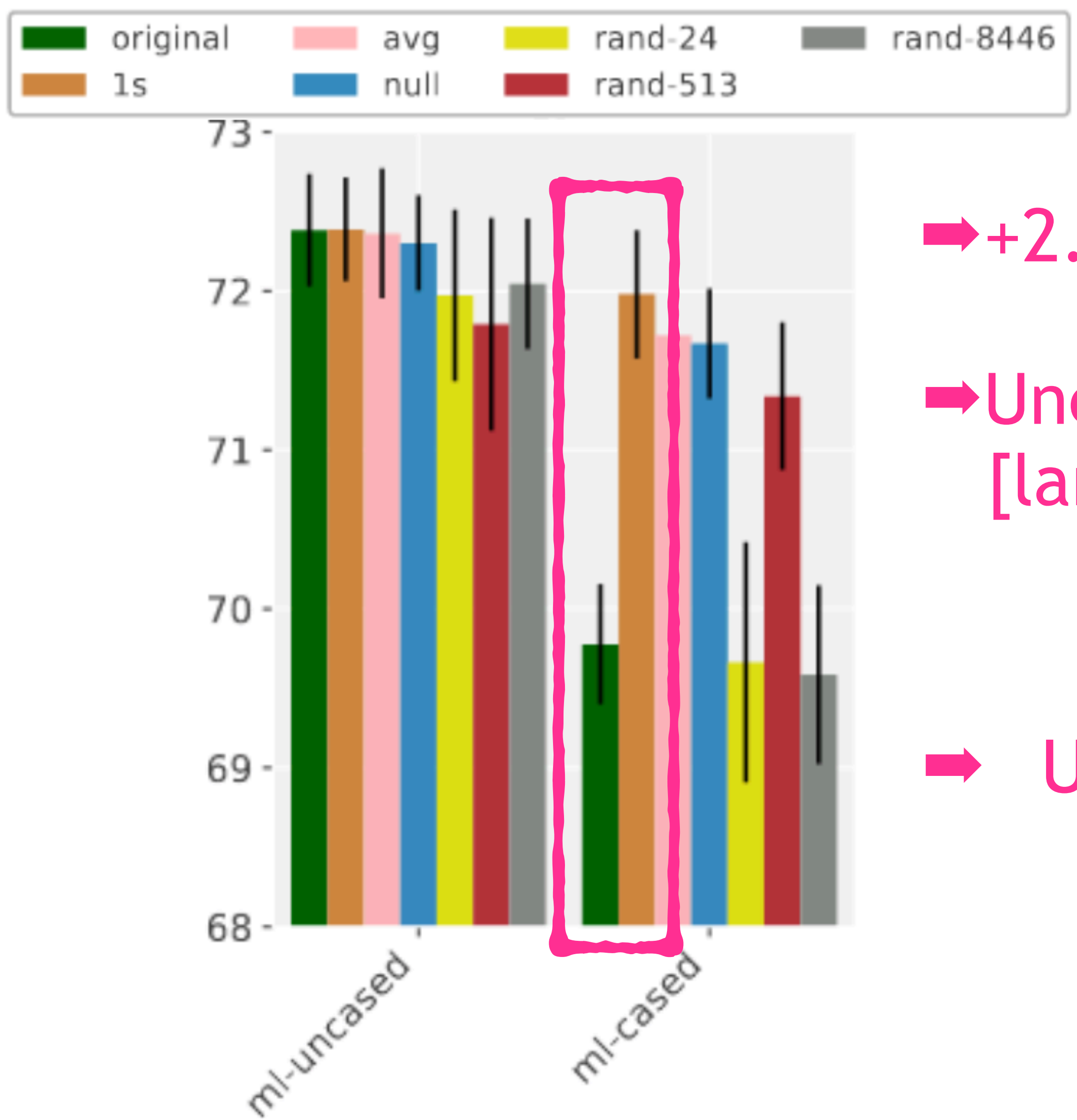
Figure 2: Visualization of the segment alternations.

Experimental Setup

- **Monolingual and Multilingual BERTs:**
  - BERT (base) cased / uncased
  - mBERT cased / mBERT uncased*
    (*not recommended according to https://github.com/google-research/bert)
- **Single-sentence prediction tasks:**
  - Sentence-level: CoLa (acceptability), SST-2 (sentiment)
  - Token-level: POS, Stemming, Morph., Dependency Parsing (similar to Udify)
- **Paired-sentence prediction tasks:**
  - GLUE tasks with paired inputs
- Additional low-resource setup (10% for UD; 1k train for other)
- Note: for LMs without NSP, segment IDs are still added during fine-tuning

Results - Largest Impact on Parsing

- Low-resource Multilingual Parsing
  (average over 9 TBs from Smith et al., 2018), 5 runs
- Large diffs for popular mBERT (cased) [trends similar for POS etc]



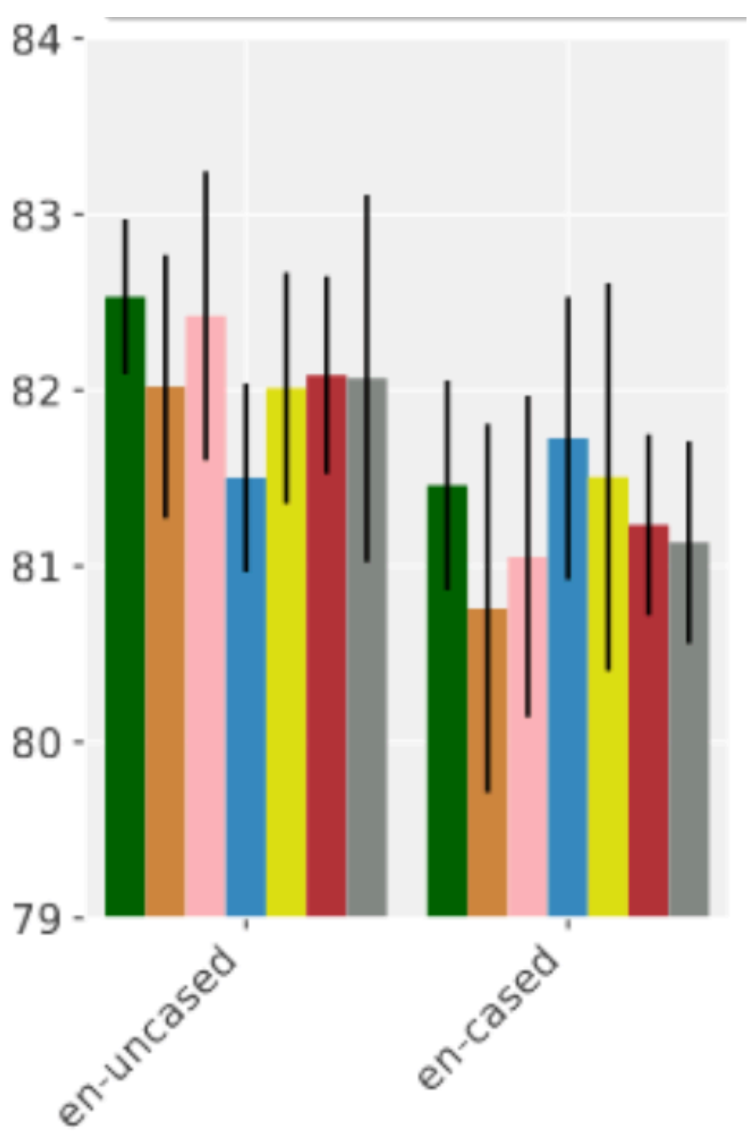
➡+2.5 LAS

➡Uncased outperforms mBERT (cased) [large due to Greek PROIEL, but not the only reason]

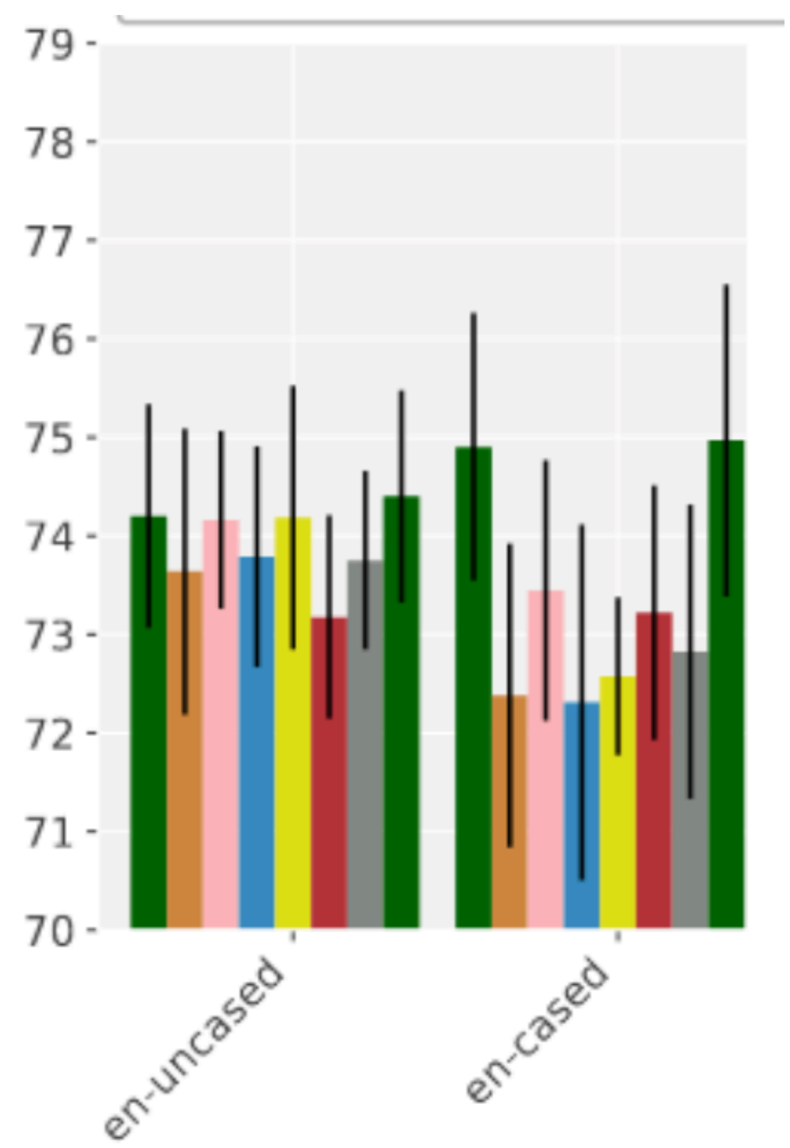
➡ Unfortunately exact pre-training differences remain unknown

Results - Sentence-level & Paired Tasks

- Close in range, despite larger fluctuations no striking difference



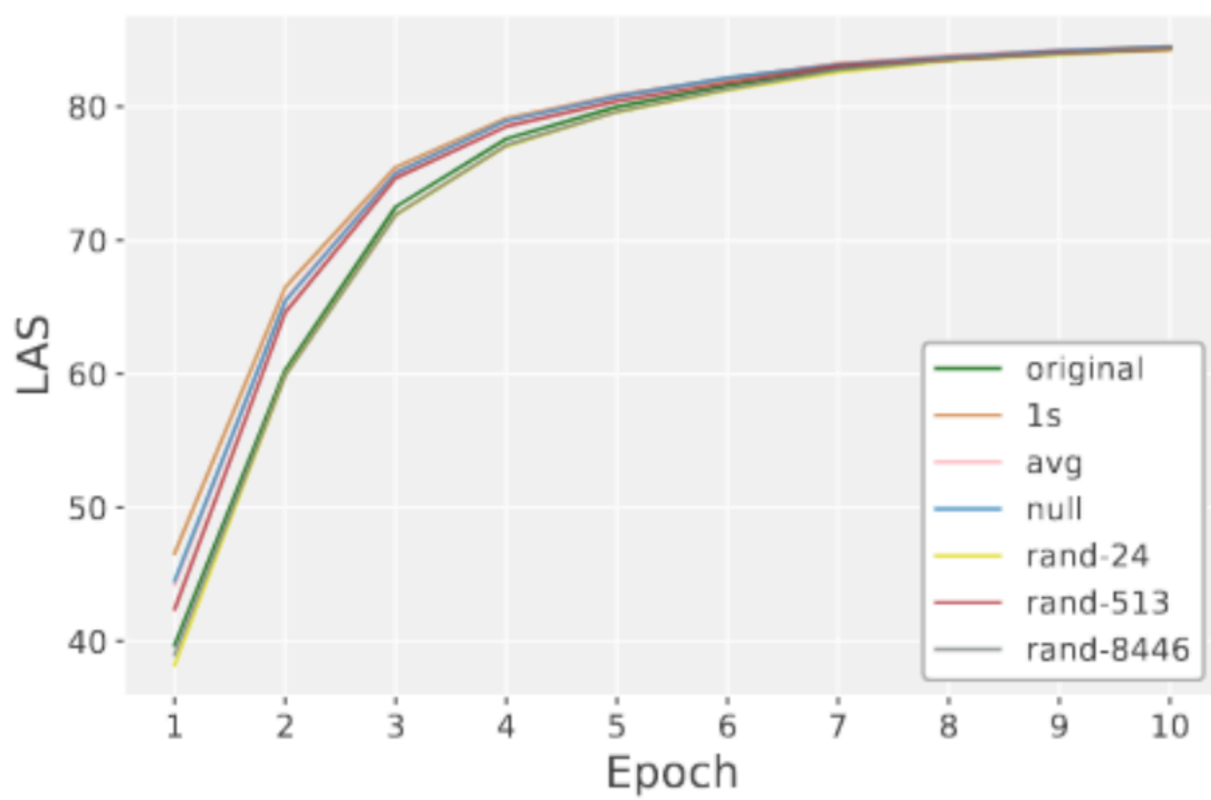
Sentence-level (CoLA, SST-2)



Sentence-paired of GLUE

What about High-Resource Parsing?

- Large diffs for mBERT (cased) disappear after 4-5 epochs
- No observable differences for high-resource multilingual parsing



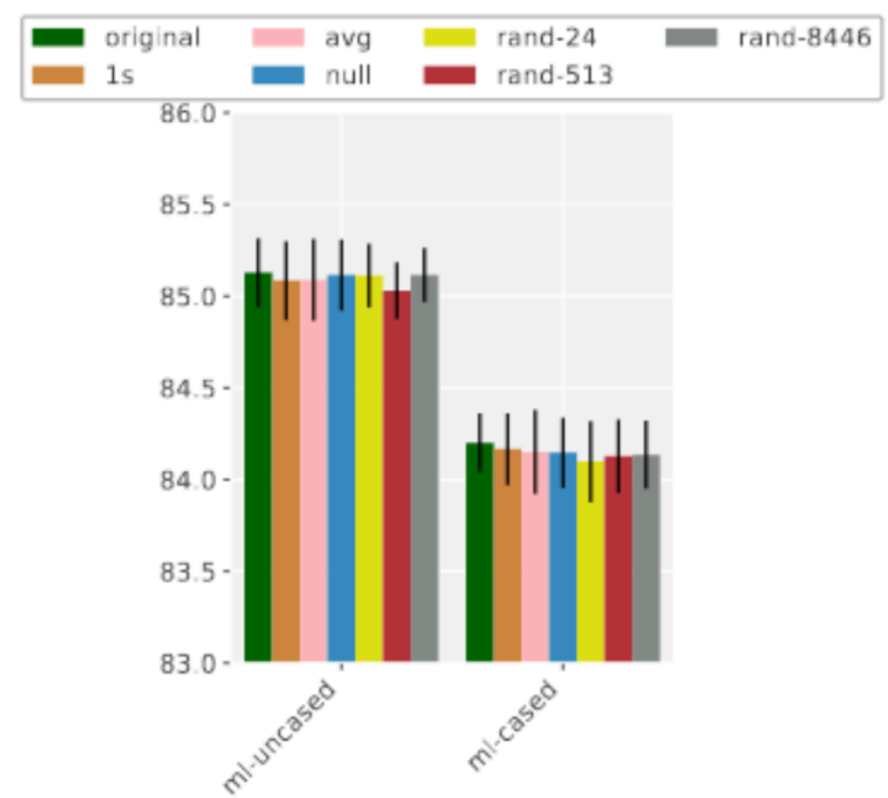
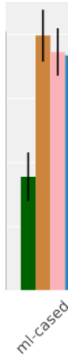
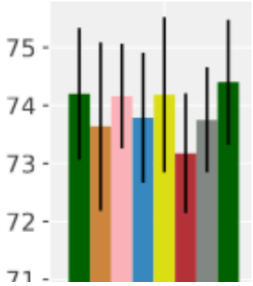


Figure 6: Average LAS scores for each setting (Section 3) on the dev data when training on full training splits. The mono-lingual embeddings are results only for EWT, the multilingual embedding results are averages over 9 treebanks.

Take-Aways



RQ1: Segment embeddings impact low-resource NLP tasks, most strikingly token-level ones



RQ2: Paired-sentence tasks and monolingual setups were impacted to modest degrees (at least for the tasks we studied)

➡ Wish: More details to be released with pre-trained language models (data, exact training setup etc)

Roadmap

1. How useful is (fortuitous) meta-data for low-res parsing?
2. How impactful are segment embeddings for low-res NLP?
3. To what extent does auxiliary data help limited training data?

MultiSkill project: Multilingual Information Extraction for Job Post Analysis

In collaboration with:

Styrelsen for Arbejdsmarked og Rekruttering

Project funded by:

DANMARKS FRIE FORSKNINGSFOND

62

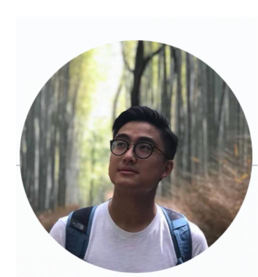
Challenges & Opportunities

- Big **multilingual** job vacancy data, on a variety of platforms
- Ultimately, can yield better job matching
    - Qs: What skills are needed? How do they change over time?
- **First step:** De-identification of personal **entities** in Job Postings, to allow sharing of data

De-identification of Privacy-related Entities in Job Postings

Kristian Nørgaard Jensen
krnj@itu.dk

Mike Zhang
mikz@itu.dk



Barbara Plank
bapl@itu.dk

Part 3

NoDaLiDa 2021

Motivation

- Most work on de-identification in the medial domain (particularly, Electronic Health Records)
  - SOTA systems mostly use LSTM-based architectures
- Personal data not only limited to that domain

JobStack

	Train	Dev	Test	Total
Time	June – August 2020	September 2020		-
# Documents	313	41	41	395
# Sentences	18,055	2082	2092	22,219
# Tokens	195,425	22,049	21,579	239,053
# Entities	4,057	462	426	5,154

- Job postings from Stackoverflow;
- Time-based data split;
- **Annotating** *Organization*, *Location*, *Profession*, *Contact*, and *Name*;
- 3 annotators.

	Token	Entity	Unlabeled
A1 – A2	0.889	0.767	0.892
A1 – A3	0.898	0.782	0.904
A2 – A3	0.917	0.823	0.920
Fleiss' κ	0.902	0.800	0.906

Annotator agreement

★ https://github.com/kris927b/JobStack

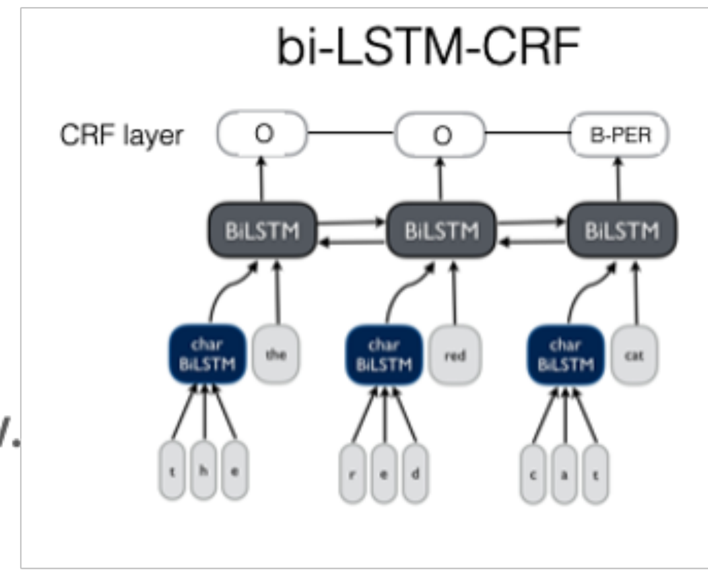
Questions

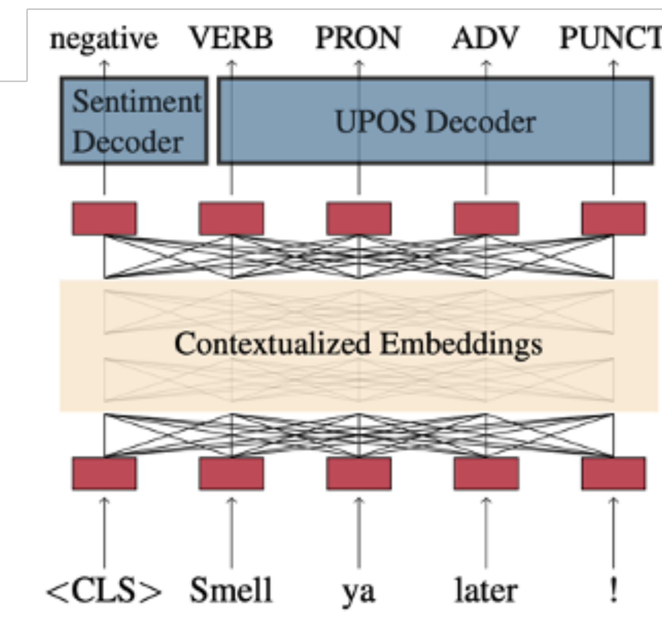
- How good is de-identification on job posting data?
- Can we leverage auxiliary data to improve performance?
  - CoNLL 2003 (NER): only some labels overlap (ORG, LOC)
  - I2b2 (EHR data): more distant genre, labels overlap more (also CONTACT, PROFESSION)

Models

- Bi-LSTM sequence tagger (*Bilty)*
    - with(out) CRF layer

- Transformer based model (MaChAmp)
    - with(out) CRF layer
    - **BERT**$_{\textbf{base}}$ (Devlin et al., 2019)
    - **BERT**$_{\textbf{overflow}}$(Tabassum et al., 2020)
        - BERT$_{\text{base}}$ architecture;
        - Q&A section of Stackoverflow.





Bilty
(Plank et al., 2016)

MaChAmp
(van der Goot et al., 2021)

Results

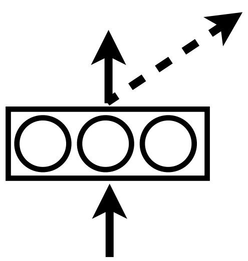
Model	Auxiliary tasks	F1 Score	Precision	Recall
Bilty + $\text{BERT}_{\text{base}}$ + CRF	JobStack	78.99 ± 0.32	$\mathbf{82.44 \pm 0.95}$	75.90 ± 1.39
MaChAmp + $\text{BERT}_{\text{base}}$ + CRF	JobStack	79.91 ± 0.38	75.92 ± 0.39	84.35 ± 0.49
	JobStack + CoNLL	81.27 ± 0.28	77.84 ± 1.19	85.06 ± 0.91
	JobStack + I2B2	$\mathbf{82.05 \pm 0.80}$	80.30 ± 0.99	83.88 ± 0.67
	JobStack + CoNLL + I2B2	81.47 ± 0.43	77.66 ± 0.58	$\mathbf{85.68 \pm 0.57}$

- I2B2 helped on PROFESSION, CoNLL on LOCATION
- Both auxiliary tasks help improve recall

Take-aways

JobStack

1. New dataset for de-identification in job postings



2. Using auxiliary data helps de-identification performance in this low-resource setup

★ Paper, Data, Code: https://arxiv.org/abs/2105.11223

★ Video (by Mike): https://www.youtube.com/watch?v=vlPQ8JAcpE0

Upcoming: Mike Zhang, Kristian Nørgaard Jensen, Sif Dam Sonniks and Barbara Plank. SkillSpan: Hard and Soft Skill Extraction from English Job Postings. In NAACL 2022.

Summary & References

1. Genre as Weak Supervision for Cross-Lingual Parsing
https://aclanthology.org/2021.emnlp-main.393/

2. A Tale on BERT and Segment Embeddings
To Appear at LREC 2022

3. De-identification of Entities in Job Postings (JobStack)
https://www.aclweb.org/anthology/W17-0200.pdf

Questions? Thanks!

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Interested?
I'm hiring PhDs
& Postdocs

@barbara_plank

B.Plank@lmu.de / bplank@itu.dk

north

IT UNIVERSITY OF COPENHAGEN

Thanks to the support by:

DANMARKS FRIE FORSKNINGSFOND

amazon

NVIDIA