

Deep-change at AXOLOTL-24: Orchestrating WSD and WSI Models for Semantic Change Modeling

Denis Kokosinskii^{1,2}, Mikhail Kuklin^{1,3}, and Nikolay Arefyev⁴

¹Moscow State University, Russia

²SaluteDevices, Russia

³Yandex, Russia

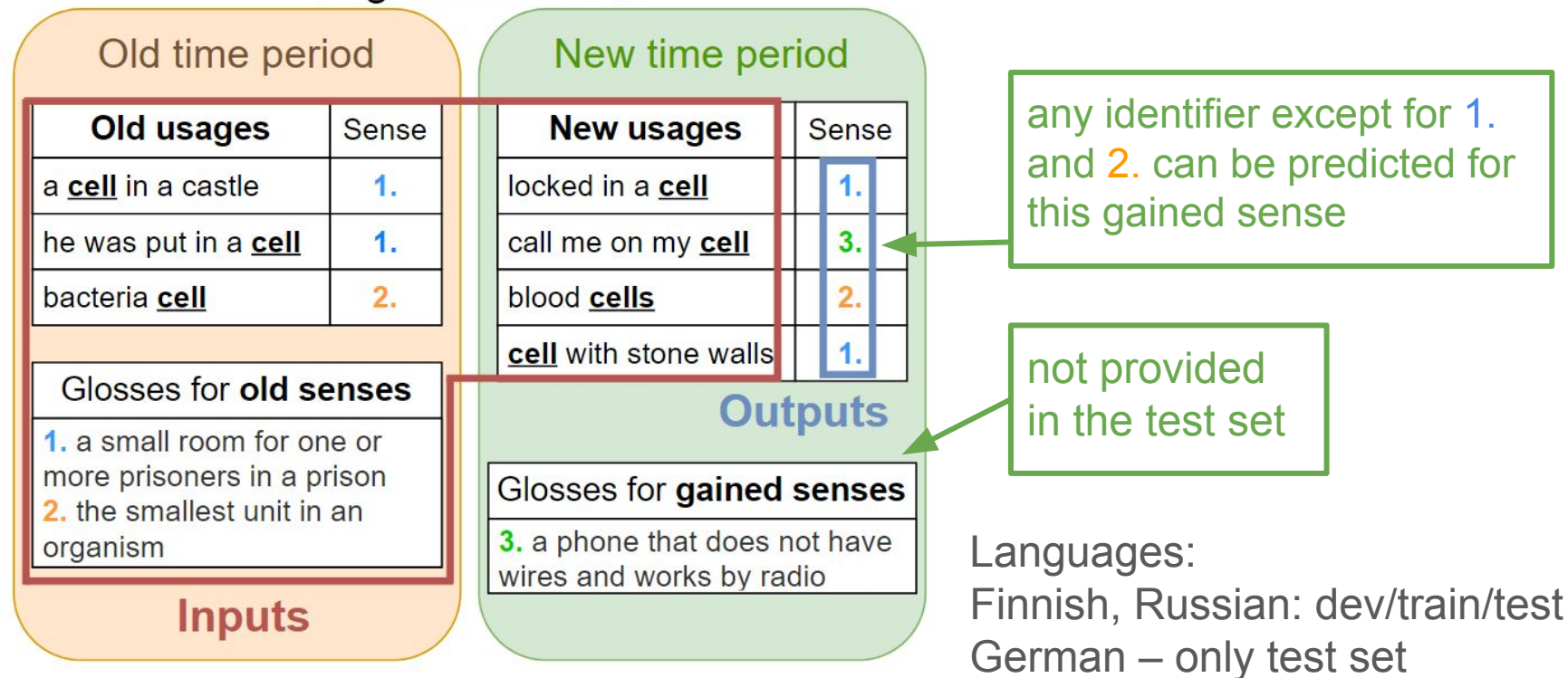
⁴University of Oslo, Norway

kokosinskiidv@my.msu.ru, kuklin.mike@yandex.ru, nikolare@uio.no

AXOLOTL-24 Subtask 1: Semantic Change Modeling (SCM)

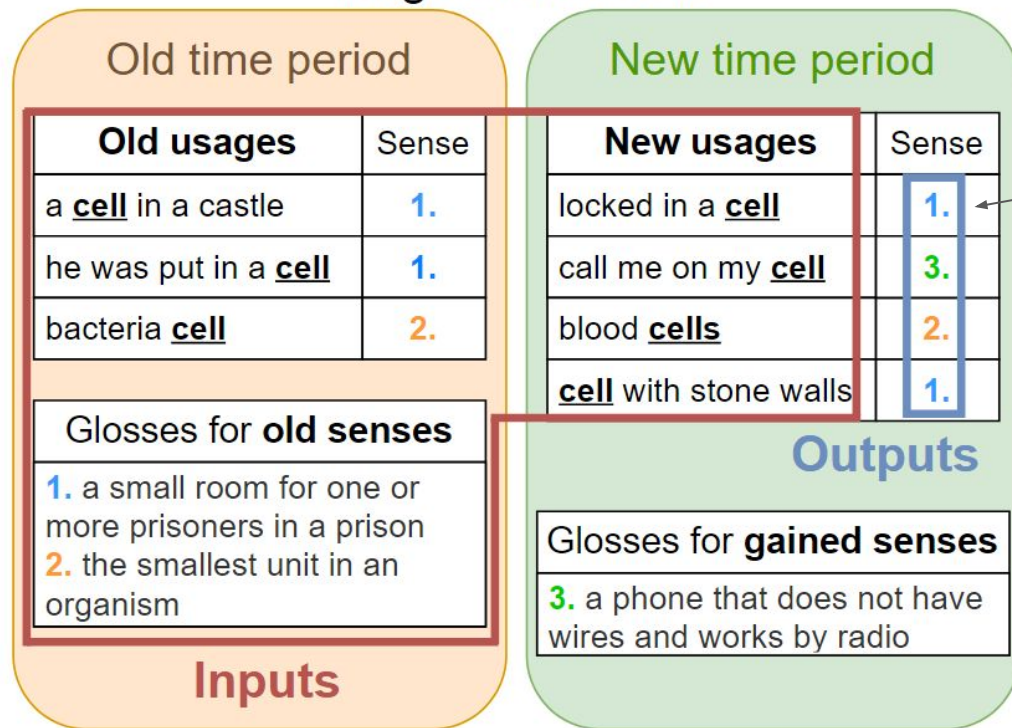
[Mariia Fedorova](#) et al. AXOLOTL'24 Shared Task on Multilingual Explainable Semantic Change Modeling, 2024

Target word: **cell**



AXOLOTL-24 Subtask 1: Metrics

Target word: **cell**



Adjusted Rand Index (ARI):
how well new usages are
grouped by meaning?

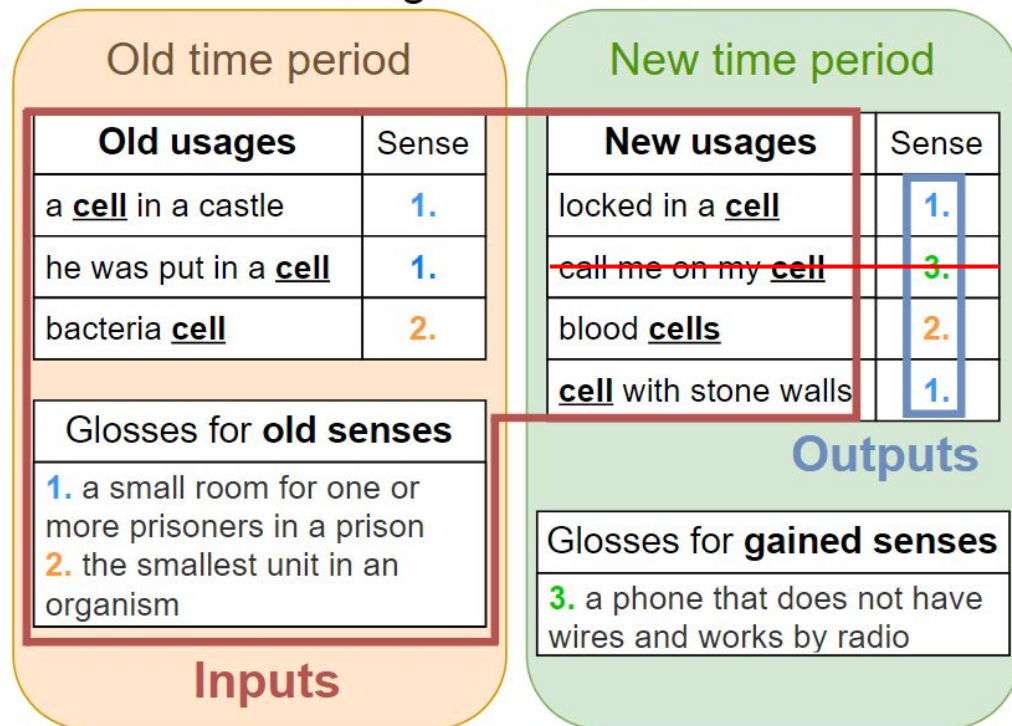
Doesn't require specific labels, e.g.
 $ARI([1,3,2,1], [a,b,c,a]) = 1.0$

RI = accuracy for pairs of usages:
same sense $\Leftrightarrow$ same cluster, diff.
senses $\Leftrightarrow$ diff. clusters

$ARI = (RI_C - Avg RI_{C'}) / (\max RI_{C'} - Avg RI_{C'})$
max and avg across clusterings C' with the same
number and sizes of clusters

AXOLOTL-24 Subtask 1: Metrics

Target word: **cell**



F1-score:

how well new usages of old senses are labeled with their senses?

2. Models

3 types of models

We employed the following types of models and their combinations:

1. **WSD: Word Sense Disambiguation models**
select among the provided (old) definitions the most suitable one ⇒
incorrectly attribute usages of gained senses to one of the old senses.
2. **WSI: Word Sense Induction models**
cluster the provided (new) usages ⇒ do not label usages with definitions
3. **SCM: Semantic Change Modeling models**
fully solve the task

WSD: the original **GlossReader (GR)** fine-tuning

GlossReader [2]: the multilingual version of BEM [1] (SOTA for WSD in English in 2020). Trained on English WSD data (SemCor), zero-shot cross-lingual transfer to other languages (XLM-R-based encoders). Employed in the winning solutions [2,3] in two LSCD shared tasks: RuShiftEval-2021 (Russian), LSCDiscovery-2022 (Spanish).

1. Blevins and Zettlemoyer. Moving Down the Long Tail of Word Sense Disambiguation with Gloss Informed Bi-encoders, 2020
2. Rachinskiy and Arefyev. Zero-shot Cross-lingual Transfer of a Gloss Language Model for Semantic Change Detection, 2021
3. Rachinskiy and Arefyev. GlossReader at LSCDiscovery: Train to Select a Proper Gloss in English – Discover Lexical Semantic Change in Spanish, 2022

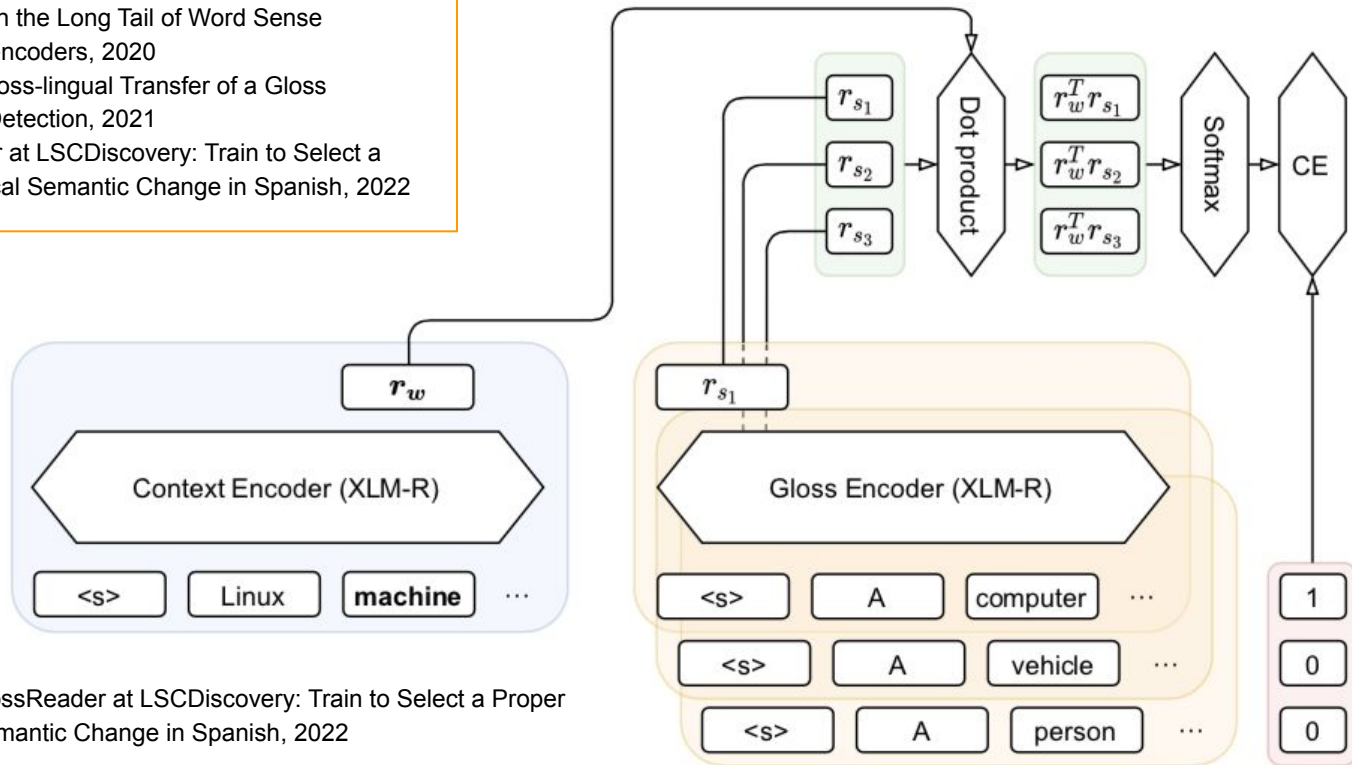
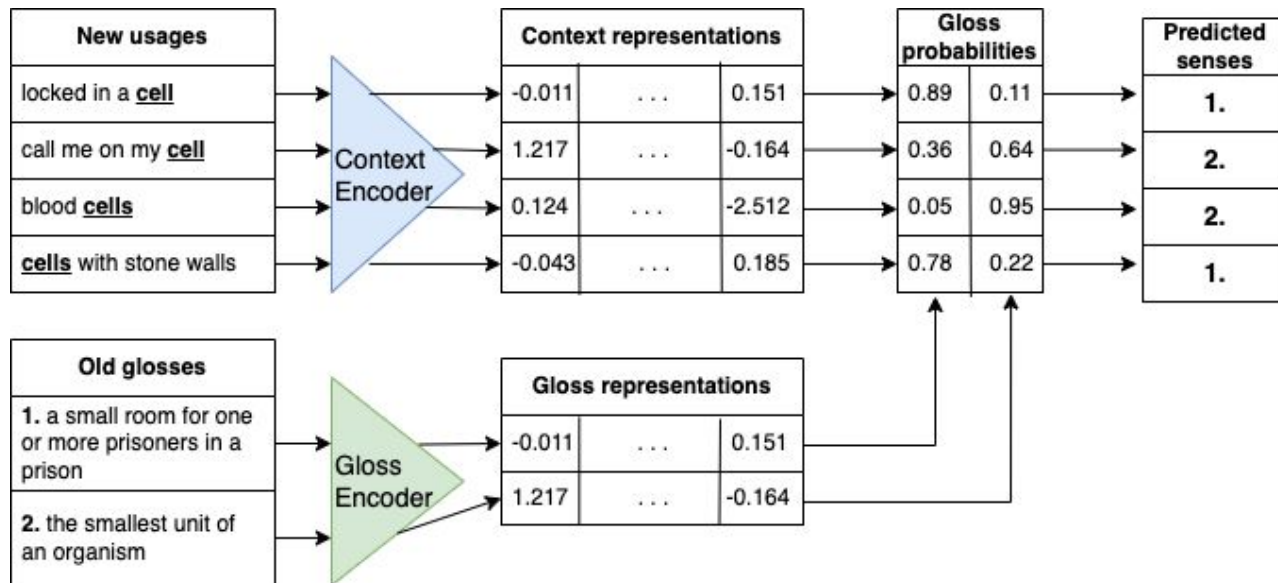


Image from: Rachinskiy & Arefyev. GlossReader at LSCDiscovery: Train to Select a Proper Gloss in English – Discover Lexical Semantic Change in Spanish, 2022

WSD: **GR** for Semantic Change Modeling

Select the most appropriate old sense for new usages

Never predicts gained senses



WSD: fine-tuning **GR Ru** and **GR FiEnRu** for AXOLOTL-24

Three datasets to further finetune GlossReader:

1. AXOLOTL-24 Russian training set
(incl. old and new usages, old and new glosses)
2. AXOLOTL-24 Finnish training set
(incl. old and new usages, old and new glosses)
3. SemCor WSD dataset in English (the original GlossReader was trained on)

- **GR Ru:** Only Russian train set is used
- **GR FiEnRu:** All three datasets are used

WSD: fine-tuning **GR Fi Special Gloss (GR Fi SG)**

Train GlossReader to detect gained senses on the Finnish train set:

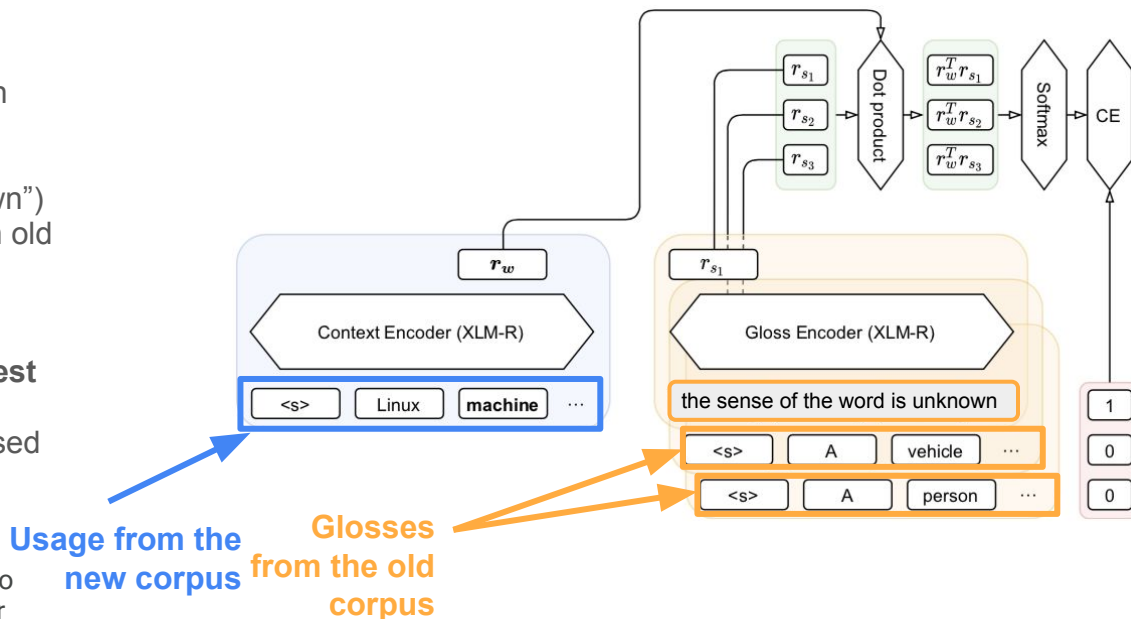
- Replace the glosses for new usages with gained senses with a **Special Gloss**
“sanan merkitystä ei tunneta”
 (“the sense of the word is unknown”)
- Include old usages and new usages with old senses as-is

Poor quality on Finnish validation set with SG*

→ **We do not use the special gloss for the test sets**

→ Simply an alternative model for WSD, not used to detect gained senses

* Possible reason: the context encoder should learn to return a context representation close to the SG if other glosses are not suitable, but it does not “see” their embeddings



WSI: Agglomerative

Inputs: new usages, N – number of clusters

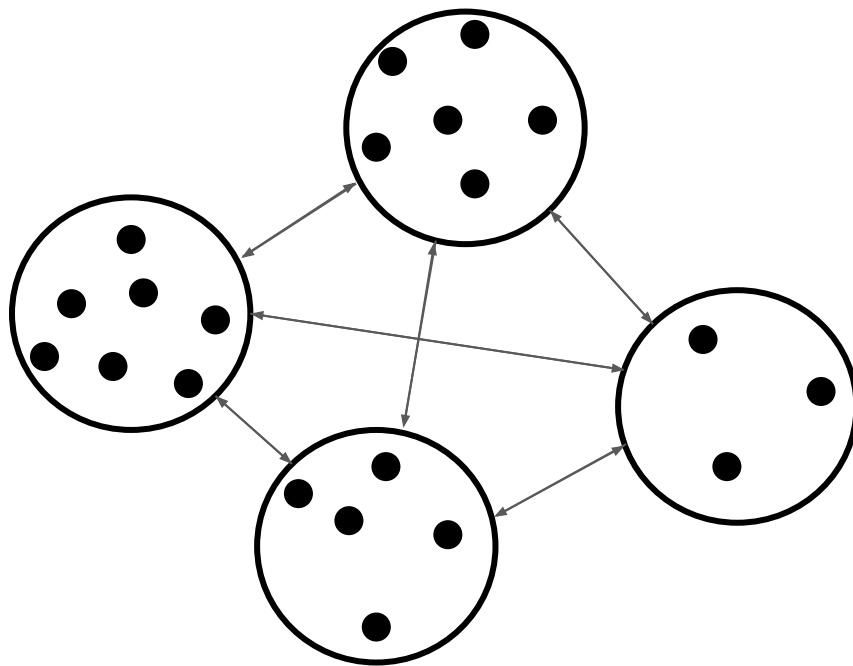
Init: each usage in its own cluster

Iterative process: on each iteration the closest pair of clusters is merged.

Stop criteria: stop after getting N clusters

Distance between usages: cosine distance between GR context embeddings.

Distance between clusters: average linkage (average distance between their usages)



● word usage, ○ cluster, ↔ compared distances

WSI: **Agglomerative**. The number of clusters

Select N between 2 and 9, maximizing the calinski-harabasz score.

For <3 usages this score is undefined
⇒ return 1 cluster

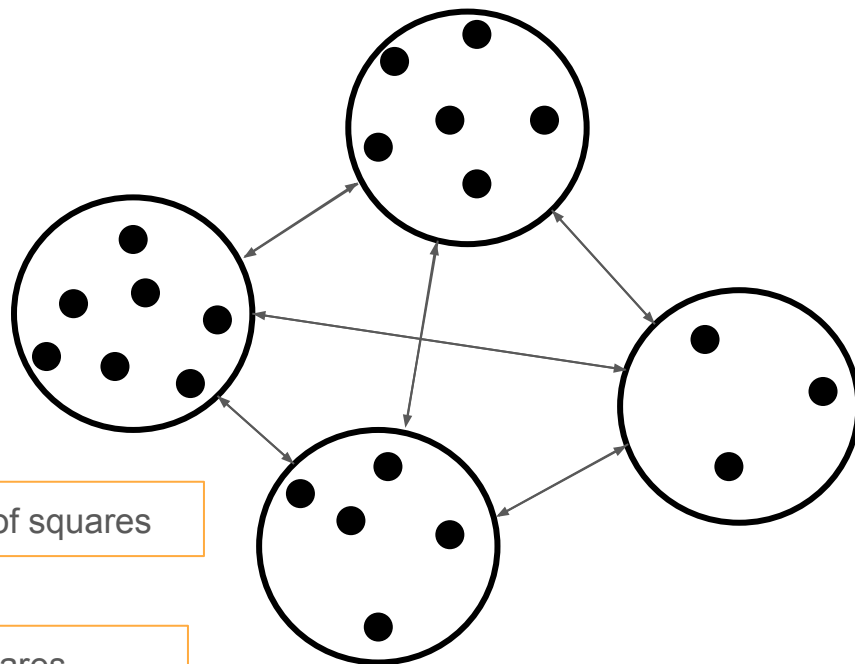
$$CH = \frac{BCSS/(k-1)}{WCSS/(n-k)}$$

$$BCSS = \sum_{i=1}^k n_i ||\mathbf{c}_i - \mathbf{c}||^2$$

between-cluster sum of squares

$$WCSS = \sum_{i=1}^k \sum_{\mathbf{x} \in C_i} ||\mathbf{x} - \mathbf{c}_i||^2$$

with-cluster sum of squares

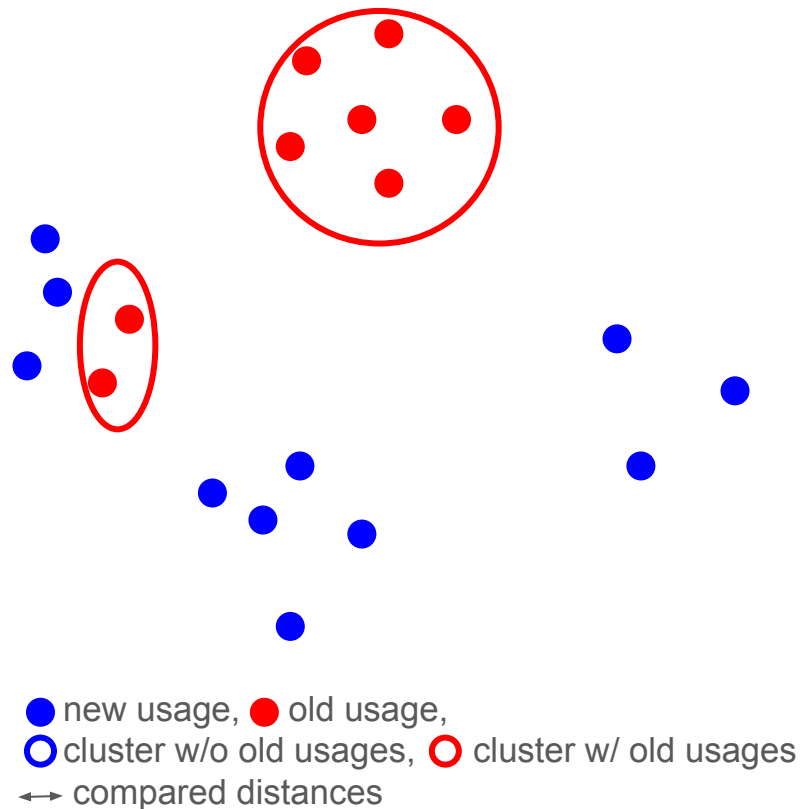


● word usage, ○ cluster, ↔ compared distances

SCM: **Agglomerative Modified (AggloM)** ← novel method

Inputs: old+new usages

Init: old usages grouped by sense, each new usage in its own cluster.



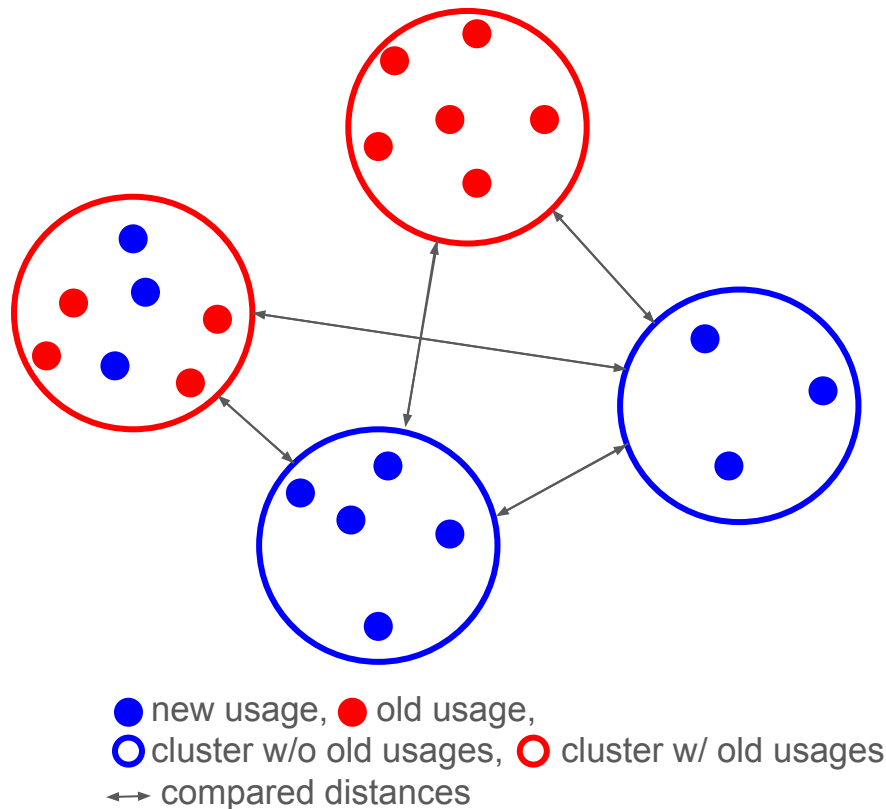
SCM: **Agglomerative Modified (AggloM)** ← novel method

Inputs: old+new usages

Init: old usages grouped by sense, each new usage in its own cluster.

Iterative process: on each iteration the closest pair of clusters one of which does not contain old usages is merged.

Stop criteria: stop after getting **#old_senses + k** clusters



SCM: **Agglomerative Modified (AggloM)** ← novel method

Inputs: old+new usages

Init: old usages grouped by sense, each new usage in its own cluster.

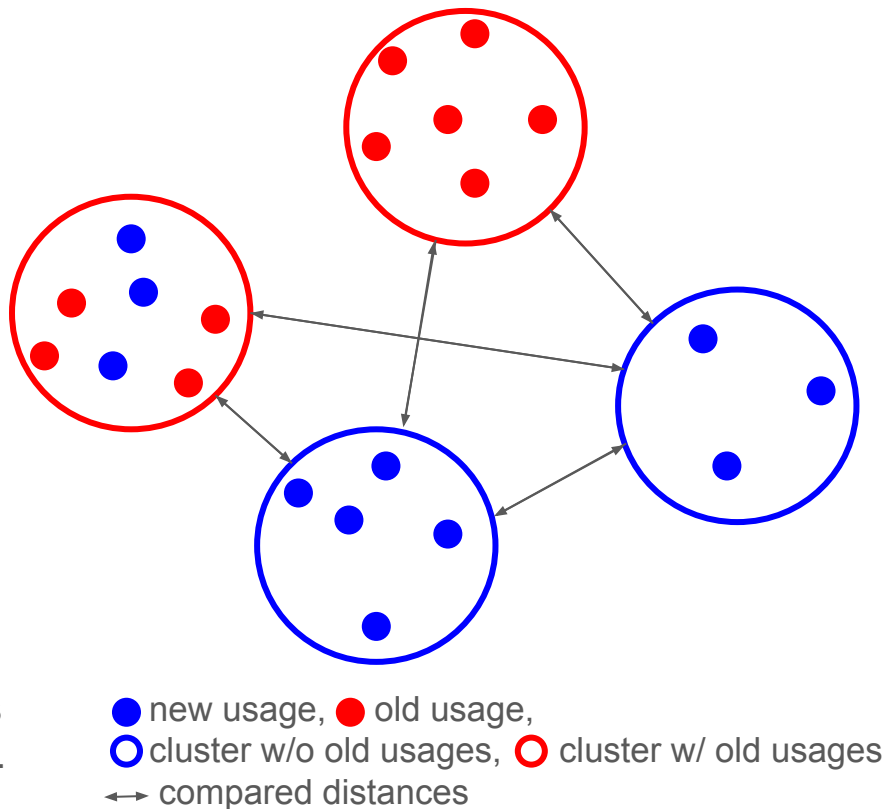
Iterative process: on each iteration the closest pair of clusters one of which does not contain old usages is merged.

Stop criteria: stop after getting **#old_senses + k** clusters

Distance between usages: cosine distance between GR context embeddings.

Distance between clusters: single linkage (minimum distance between their usages).

On Fi dev we selected $k=0 \Rightarrow$ each new usages ends in a cluster corresponding to an old sense.
Did not use for Ru (often no old usages)



SCM: **Cluster2Sense** ← novel method

Inputs:

set $C1$ of WSI predicted clusters;
set $C2$ of WSD predicted clusters
(corresponding to old senses);

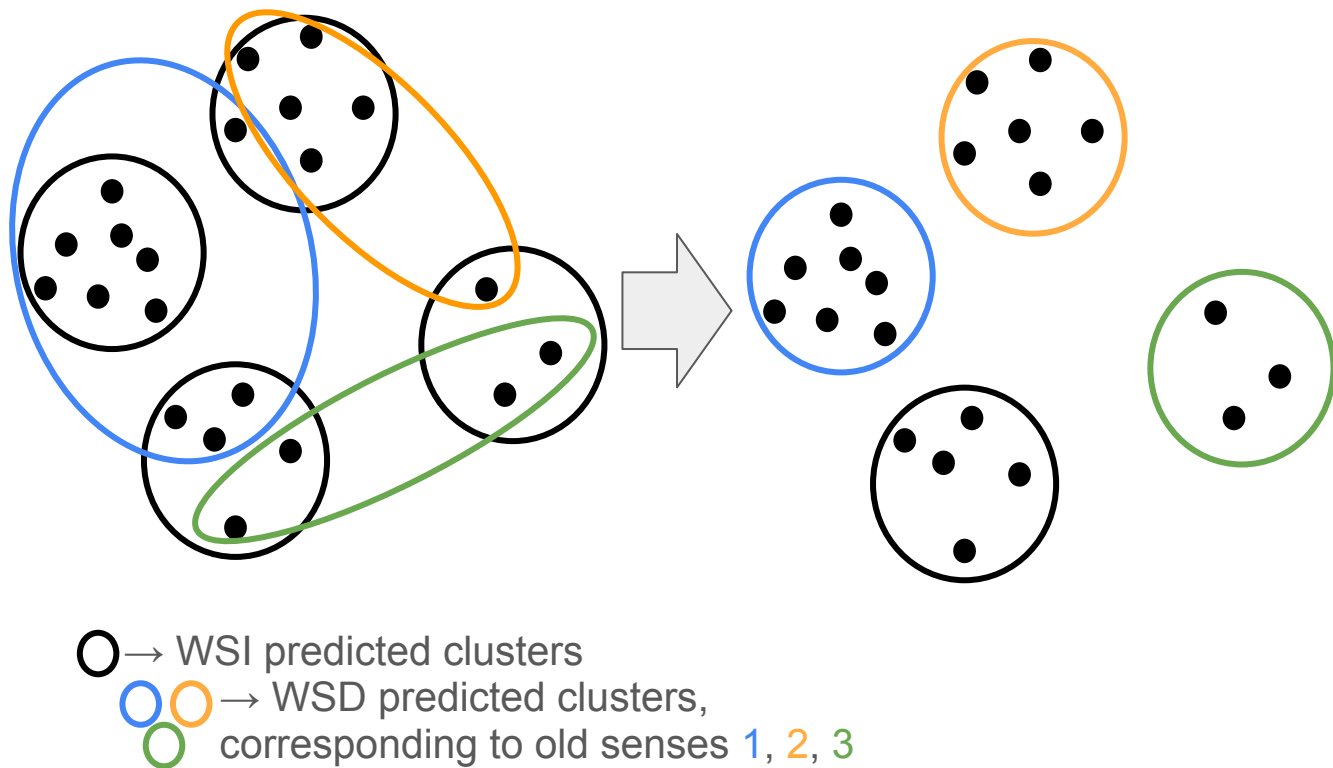
Iterative process:

- 1) select a pair of clusters $\{(c1, c2) \mid c1 \in C1; c2 \in C2\}$ with the highest Jaccard similarity.
- 2) Relabel $c1$ as $c2$ (old sense).
- 3) Remove $c1$ from $C1$, $c2$ from $C2$.

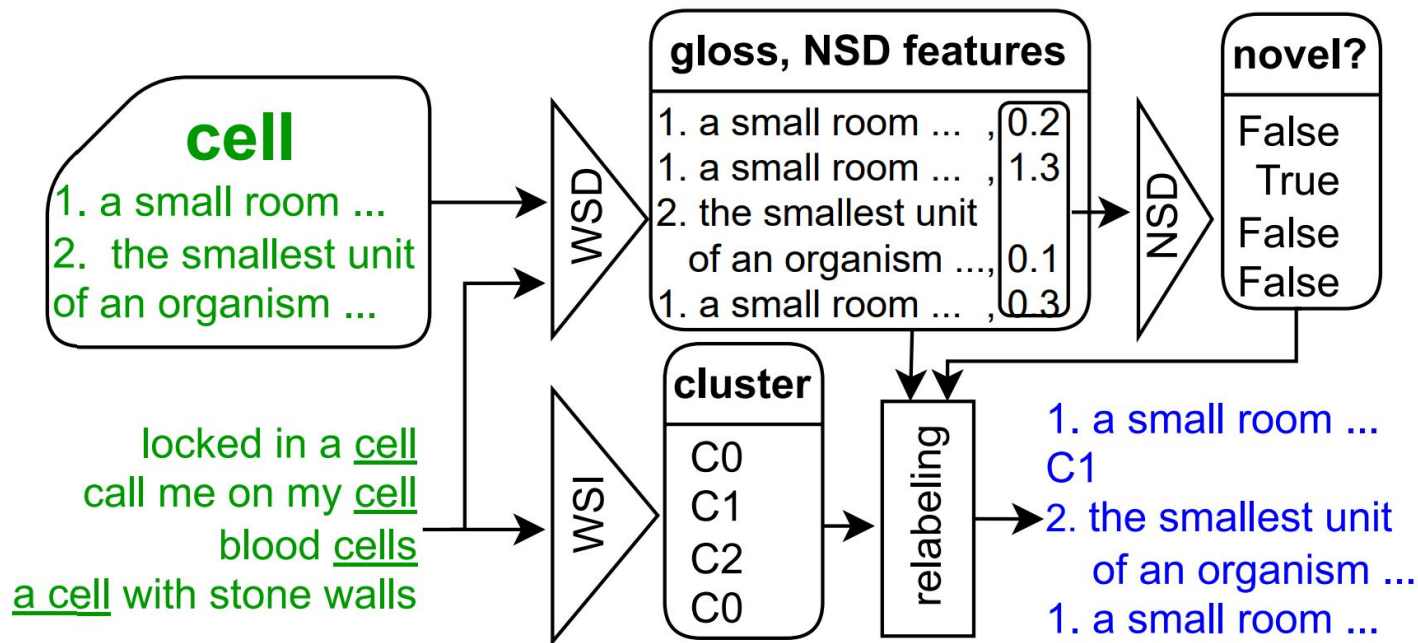
Stop criteria: $C1$ or $C2$ is empty

Case $|C1| > |C2|$: some clusters are not relabeled → gained senses

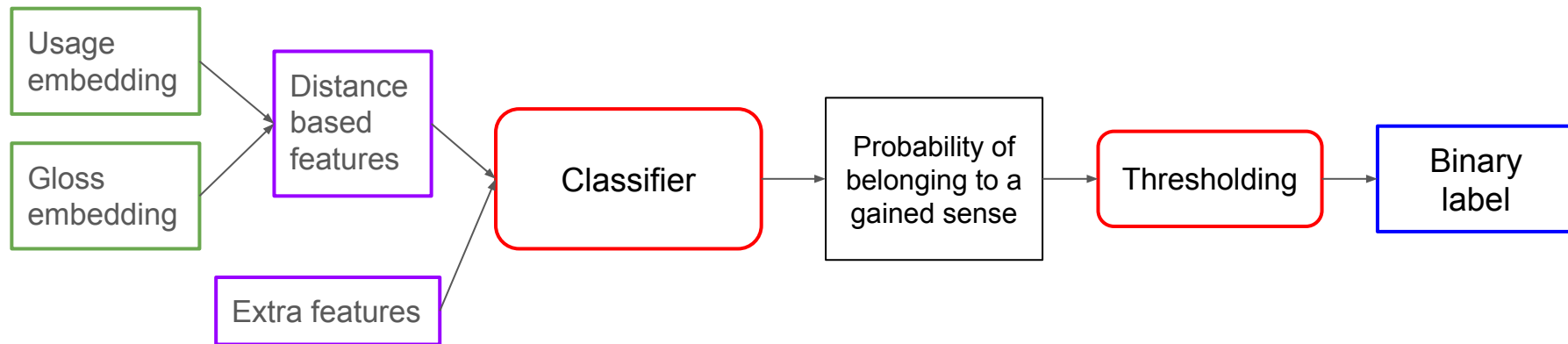
Case $|C1| < |C2|$: some old senses will not map to clusters → lost senses



SCM: **Outlier2Cluster** ← novel method



NSD: Architecture



Gloss embedding - embedding of a gloss selected by the WSD model for this usage.

Classifier - logistic regression with L2 regularization of $C=1$. It was trained on AXOLOTL Russian/Finnish dev set. For German we tried both Russian and Finnish dev sets.

Binary label - output of NSD model, whether it is a usage of a gained sense.

NSD: Features

For NSD mostly distance based features were used.

Additionally we try extra features:

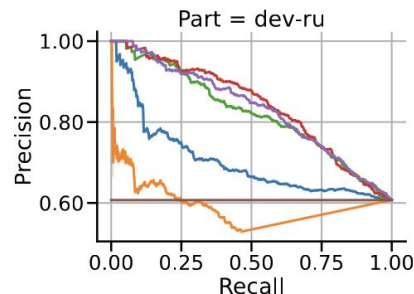
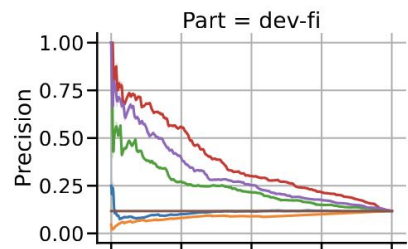
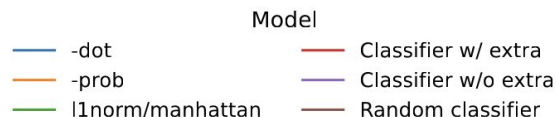
- #old senses
- #old usages
- #new usages

We employ the Standard Scaler to normalize features – poor performance otherwise.

Distance Function		Cos.	Euclid.	Manh.
Encoders	Normalized			
GR FiEnRu	No	✓	✓	✓
GR FiEnRu	L1-norm			✓
GR FiEnRu	L2-norm		✓	
GR	No	✓	✓	✓
GR	L1-norm			✓
GR	L2-norm		✓	

Ten distance-based features used in the NSD model. Distances are calculated between usage and gloss representations obtained from context and gloss encoders of the same GlossReader model. GR stand for GlossReader, Cos. is the cosine distance, Euclid. is the euclidean distance, Manh. is the manhattan distance.

NSD model ablation study. Precision-recall curves & AP

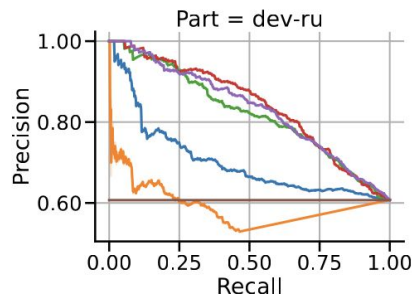
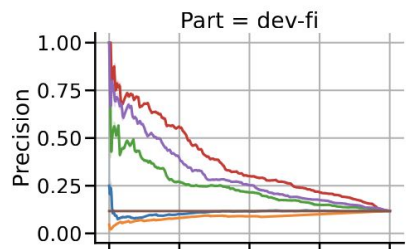
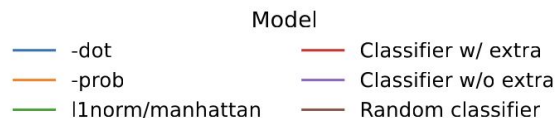


Model	dev fi AP		dev ru AP	
	GR	GR FiEnRu	GR	GR FiEnRu
single features				
cosine	0.106	0.110	0.685	0.695
euclid.	0.106	0.110	0.684	0.694
l2/euclid.	0.106	0.110	0.685	0.695
manh.	0.106	0.113	0.685	0.690
l1/manh.	0.154	0.242	0.816	0.822
full classifiers				
classifier w/ extra	0.378		0.840	
classifier w/o extra	0.305		0.833	
best pairs of features w/o extra features				
l1/manh. + euclid.	0.194	0.284	0.818	0.823
l1/manh. + l2/euclid.	0.195	0.284	0.818	0.823
l1/manh. + manh.	0.192	0.277	0.819	0.823
best pairs of features w/ extra features				
l1/manh. + #old usages	0.190	0.291	0.820	0.827
l1/manh. + #new usages	0.153	0.249	0.821	0.829
#new usages + #old senses	0.266	0.266	0.643	0.643

Underlined - the best of all
Bold - the best in group

the predicted probability of the selected gloss and dot product perform much worse than a classifier. On Finnish they are comparable to a random classifier.

NSD model ablation study. Best distance

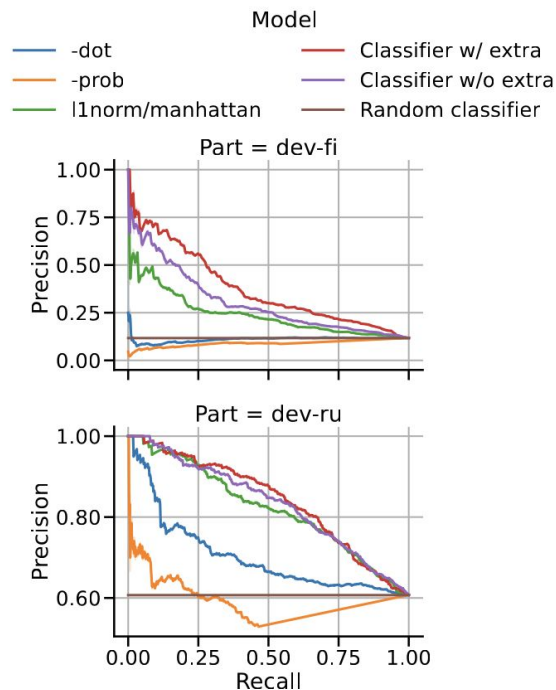


Model	dev fi AP		dev ru AP	
	GR	GR FiEnRu	GR	GR FiEnRu
single features				
cosine	0.106	0.110	0.685	0.695
euclid.	0.106	0.110	0.684	0.694
l2/euclid.	0.106	0.110	0.685	0.695
manh.	0.106	0.113	0.685	0.690
l1/manh.	0.154	0.242	0.816	0.822
full classifiers				
classifier w/ extra		0.378		0.840
classifier w/o extra		0.305		0.833
best pairs of features w/o extra features				
l1/manh. + euclid.	0.194	0.284	0.818	0.823
l1/manh. + l2/euclid.	0.195	0.284	0.818	0.823
l1/manh. + manh.	0.192	0.277	0.819	0.823
best pairs of features w/ extra features				
l1/manh. + #old usages	0.190	0.291	0.820	0.827
l1/manh. + #new usages	0.153	0.249	0.821	0.829
#new usages + #old senses	0.266	0.266	0.643	0.643

l1/manh.
performs way
better than
other distances

Underlined - the best of all
Bold - the best in group

NSD model ablation study. GR finetuning on AXOLOTL data

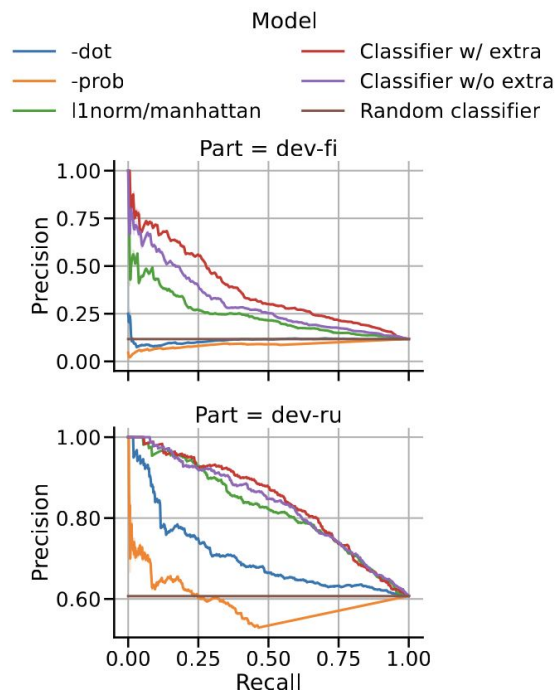


Model	dev fi AP		dev ru AP	
	GR	GR FiEnRu	GR	GR FiEnRu
single features				
cosine	0.106	0.110	0.685	0.695
euclid.	0.106	0.110	0.684	0.694
l2/euclid.	0.106	0.110	0.685	0.695
manh.	0.106	0.113	0.685	0.690
l1/manh.	0.154	0.242	0.816	0.822
full classifiers				
classifier w/ extra	0.378		0.840	
classifier w/o extra	0.305		0.833	
best pairs of features w/o extra features				
l1/manh. + euclid.	0.194	0.284	0.818	0.823
l1/manh. + l2/euclid.	0.195	0.284	0.818	0.823
l1/manh. + manh.	0.192	0.277	0.819	0.823
best pairs of features w/ extra features				
l1/manh. + #old usages	0.190	0.291	0.820	0.827
l1/manh. + #new usages	0.153	0.249	0.821	0.829
#new usages + #old senses	0.266	0.266	0.643	0.643

for the Finish dataset
GlossReader provides
poor embedding
without fine-tuning

Underlined - the best of all
Bold - the best in group

NSD model ablation study. Extra features



Model	dev fi AP		dev ru AP	
	GR	GR FiEnRu	GR	GR FiEnRu
single features				
cosine	0.106	0.110	0.685	0.695
euclid.	0.106	0.110	0.684	0.694
l2/euclid.	0.106	0.110	0.685	0.695
manh.	0.106	0.113	0.685	0.690
l1/manh.	0.154	0.242	0.816	0.822
full classifiers				
classifier w/ extra	0.378		0.840	
classifier w/o extra	0.305		0.833	
best pairs of features w/o extra features				
l1/manh. + euclid.	0.194	0.284	0.818	0.823
l1/manh. + l2/euclid.	0.195	0.284	0.818	0.823
l1/manh. + manh.	0.192	0.277	0.819	0.823
best pairs of features w/ extra features				
l1/manh. + #old usages	0.190	0.291	0.820	0.827
l1/manh. + #new usages	0.153	0.249	0.821	0.829
#new usages + #old senses	0.266	0.266	0.643	0.643

extra features improve NSD model, especially on Finnish

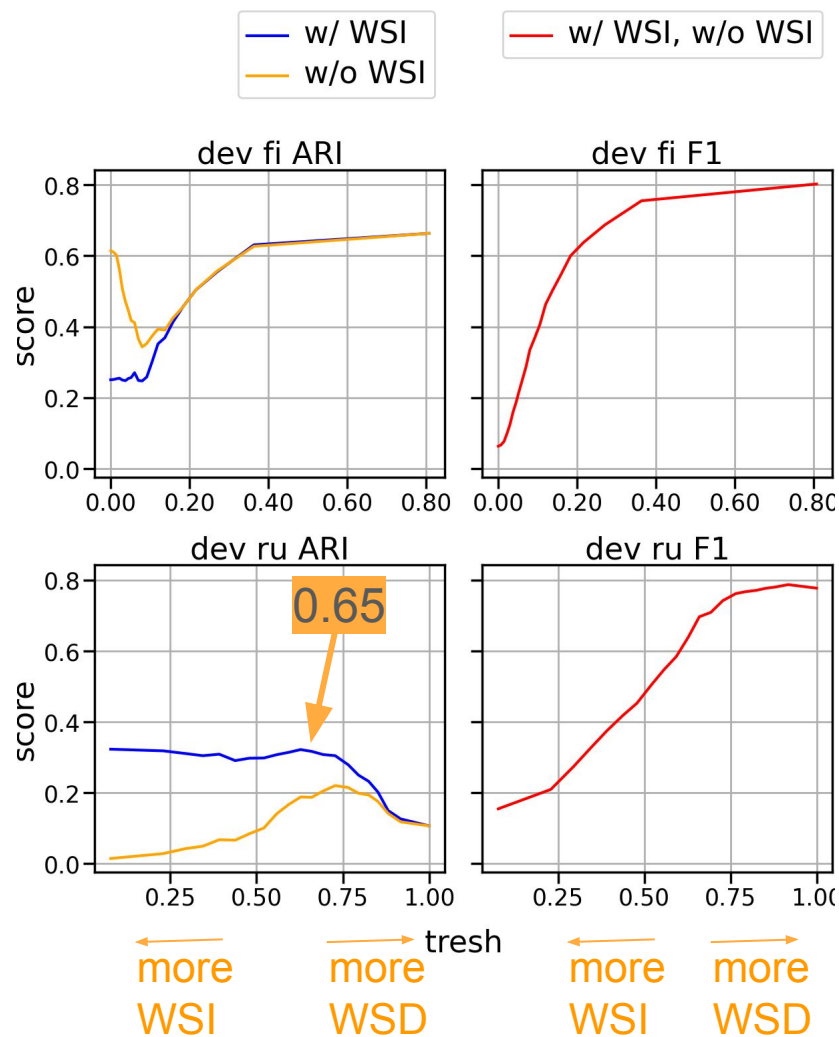
Underlined - the best of all
Bold - the best in group

NSD: Threshold Selection

Threshold of 0.65, selected on dev sets
(mostly Russian)

Russian: near-optimal ARI, small loss in F1,
WSI labels for ~42% of usages

Finnish: rarely returns WSI labels (<1%),
almost like pure WSD



3. Test results

SOTA

SOTA

SOTA

Method	ARI					F1				
	Fi	Ru	De	FiRu	AVG	Fi	Ru	De	FiRu	AVG
WSD methods										
GR	0.581	0.041	0.386	0.311	0.336	0.690	◇0.721	0.694	0.706	0.702
GR FiEnRu	◇ 0.649	0.048	◇0.521	0.348	0.406	◇ 0.756	◇ 0.750	◇0.745	◇ 0.753	◇ 0.750
GR Ru	0.568	0.053	0.464	0.310	0.361	0.568	◇ 0.750	◇0.724	0.659	0.681
GR Fi SG	◇0.638	0.059	◇ 0.543	0.348	0.413	◇0.752	◇0.729	◇ 0.758	◇0.741	◇0.746
WSI methods										
Agglomerative	0.209	◇ 0.259	0.316	0.234	0.261	0.055	0.152	0.042	0.104	0.083
SCM methods										
AggloM	0.581	0	0.492	0.290	0.357	0.674	0	0.695	0.337	0.456
AggloM FiEnRu	◇0.631	0	0.485	0.315	0.372	◇0.731	0	0.639	0.366	0.457
Cluster2Sense	0.209	◇ 0.259	0.316	0.234	0.261	0.392	0.346	0.432	0.369	0.390
Outlier2Cluster ^{ru} _{fi}	◇ 0.649	◇0.247	^{0.322} _{0.480}	◇ 0.448	^{0.406} _{◇0.459}	◇ 0.756	0.645	^{0.510} _{◇0.745}	0.701	^{0.637} _{◇0.715}
Other teams										
Holotniekat	0.596	0.043	0.298	0.319	0.312	0.655	0.661	0.608	0.658	0.641
TartuNLP	0.437	0.098	0.396	0.267	0.310	0.550	0.640	0.580	0.595	0.590
IMS_Stuttgart	0.548	0	0.314	0.274	0.287	0.590	0.570	0.300	0.580	0.487
ABDN-NLP	0.553	0.009	0.102	0.281	0.221	0.655	0	0.638	0.328	0.431
WooperNLP	0.428	0.132	0	0.280	0.186	0.503	0.446	0	0.475	0.316
Baseline	0.023	0.079	0.022	0.051	0.041	0.230	0.260	0.130	0.245	0.207

Underlined - the best of all

Bold - the best in group

◇ - within 5% (not percentage points) from the best

SOTA - outperforms all other participants according to all official leaderboard metrics (ARI and F1 averaged across all languages and FiRu)

>baselines
>baselines
>baselines
>baselines

>baselines

>baselines

- outperforms all baselines according to all official leaderboard metrics (ARI and F1 averaged across all languages and FiRu)
(NB: AggloM is not applicable for Russian)

Method	ARI					F1				
	Fi	Ru	De	FiRu	AVG	Fi	Ru	De	FiRu	AVG
WSD methods										
GR	0.581	0.041	0.386	0.311	0.336	0.690	◇0.721	0.694	0.706	0.702
GR FiEnRu	◇ 0.649	0.048	◇0.521	0.348	0.406	◇ 0.756	◇ 0.750	◇0.745	◇ 0.753	◇ 0.750
GR Ru	0.568	0.053	0.464	0.310	0.361	0.568	◇ 0.750	◇0.724	0.659	0.681
GR Fi SG	◇0.638	0.059	◇ 0.543	0.348	0.413	◇0.752	◇0.729	◇ 0.758	◇0.741	◇0.746
WSI methods										
Agglomerative	0.209	◇ 0.259	0.316	0.234	0.261	0.055	0.152	0.042	0.104	0.083
SCM methods										
AggloM	0.581	0	0.492	0.290	0.357	0.674	0	0.695	0.337	0.456
AggloM FiEnRu	◇0.631	0	0.485	0.315	0.372	◇0.731	0	0.639	0.366	0.457
Cluster2Sense	0.209	◇ 0.259	0.316	0.234	0.261	0.392	0.346	0.432	0.369	0.390
Outlier2Cluster ^{ru} _{fi}	◇ 0.649	◇0.247	^{0.322} _{0.480}	◇ 0.448	^{0.406} _{◇0.459}	◇ 0.756	0.645	^{0.510} _{◇0.745}	0.701	^{0.637} _{◇0.715}
Other teams										
Holotniekat	0.596	0.043	0.298	0.319	0.312	0.655	0.661	0.608	0.658	0.641
TartuNLP	0.437	0.098	0.396	0.267	0.310	0.550	0.640	0.580	0.595	0.590
IMS_Stuttgart	0.548	0	0.314	0.274	0.287	0.590	0.570	0.300	0.580	0.487
ABDN-NLP	0.553	0.009	0.102	0.281	0.221	0.655	0	0.638	0.328	0.431
WooperNLP	0.428	0.132	0	0.280	0.186	0.503	0.446	0	0.475	0.316
Baseline	0.023	0.079	0.022	0.051	0.041	0.230	0.260	0.130	0.245	0.207
Random	0.522 ±0.007	0.004 ±0.010	0.042 ±0.011	0.264	0.190	0.621 ±0.007	0.577 ±0.012	0.399 ±0.019	0.599	0.532
MFS for old usages	0.596	0.005	0.125	0.300	0.242	0.654 ±0.005	0.577 ±0.015	0.344 ±0.021	0.616	0.525

4. Discussion

ARI: how well new usages are grouped by their meaning?

WSI gives much better grouping than WSD for Ru, but much worse for Fi

Method	Fi	Ru	ARI De	FiRu	AVG
WSD methods					
GR	0.581	0.041	0.386	0.311	0.336
GR FiEnRu	<u>◇0.649</u>	0.048	◇0.521	0.348	0.406
GR Ru	0.568	0.053	0.464	0.310	0.361
GR Fi SG	◇0.638	0.059	◇0.543	0.348	0.413
WSI methods					
Agglomerative	0.209	◇0.259	0.316	0.234	0.261
SCM methods					
AggloM	0.581	0	0.492	0.290	0.357
AggloM FiEnRu	◇0.631	0	0.485	0.315	0.372
Cluster2Sense	0.209	◇0.259	0.316	0.234	0.261
Outlier2Cluster _{fi} ^{ru}	◇0.649	◇0.247	0.322 0.480	◇0.448	0.406 ◇0.459

Underlined - the best of all

Bold - the best in group

◇ - within 5% (not percentage points) from the best

ARI: how well new usages are grouped by their meaning?

WSI gives much better grouping than WSD for Ru, but much worse for Fi and a bit worse for De

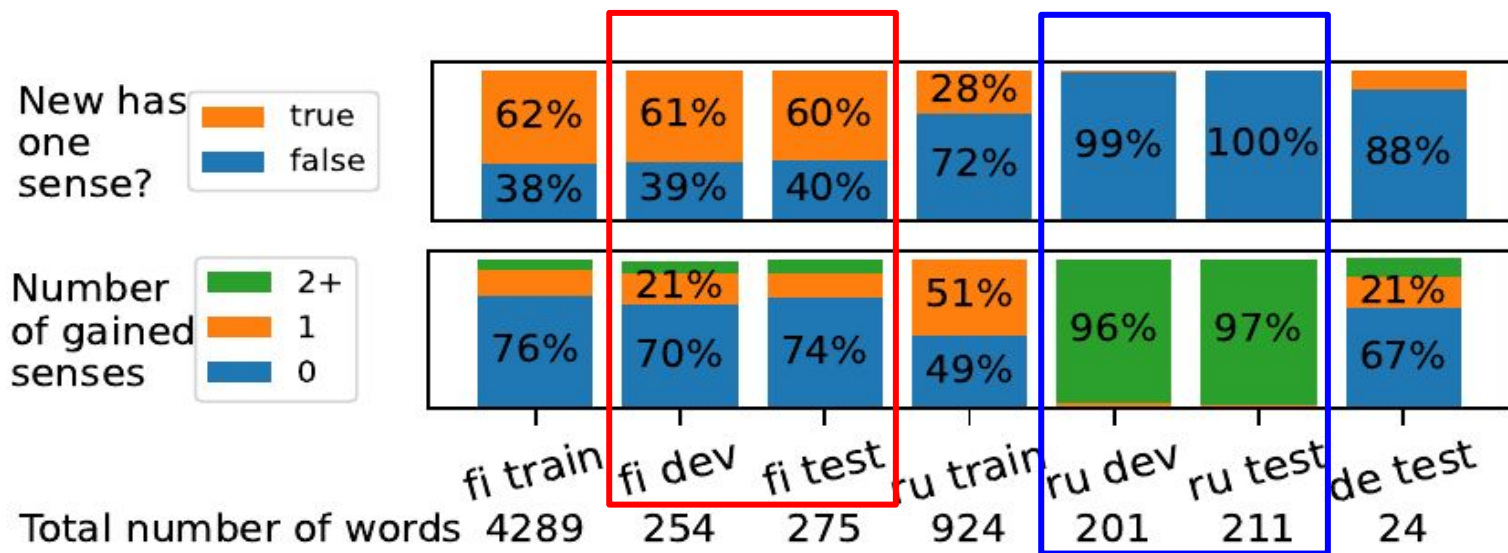
Method	Fi	Ru	ARI De	FiRu	AVG
WSD methods					
GR	0.581	0.041	0.386	0.311	0.336
GR FiEnRu	<u>◇0.649</u>	0.048	◇0.521	0.348	0.406
GR Ru	0.568	0.053	0.464	0.310	0.361
GR Fi SG	◇0.638	0.059	<u>◇0.543</u>	0.348	0.413
WSI methods					
Agglomerative	0.209	<u>◇0.259</u>	0.316	0.234	0.261
SCM methods					
AggloM	0.581	0	0.492	0.290	0.357
AggloM FiEnRu	◇0.631	0	0.485	0.315	0.372
Cluster2Sense	0.209	<u>◇0.259</u>	0.316	0.234	0.261
Outlier2Cluster _{fi} ^{ru}	<u>◇0.649</u>	◇0.247	<i>0.322</i> <i>0.480</i>	<u>◇0.448</u>	<i>0.406</i> <u>◇0.459</u>

Underlined - the best of all

Bold - the best in group

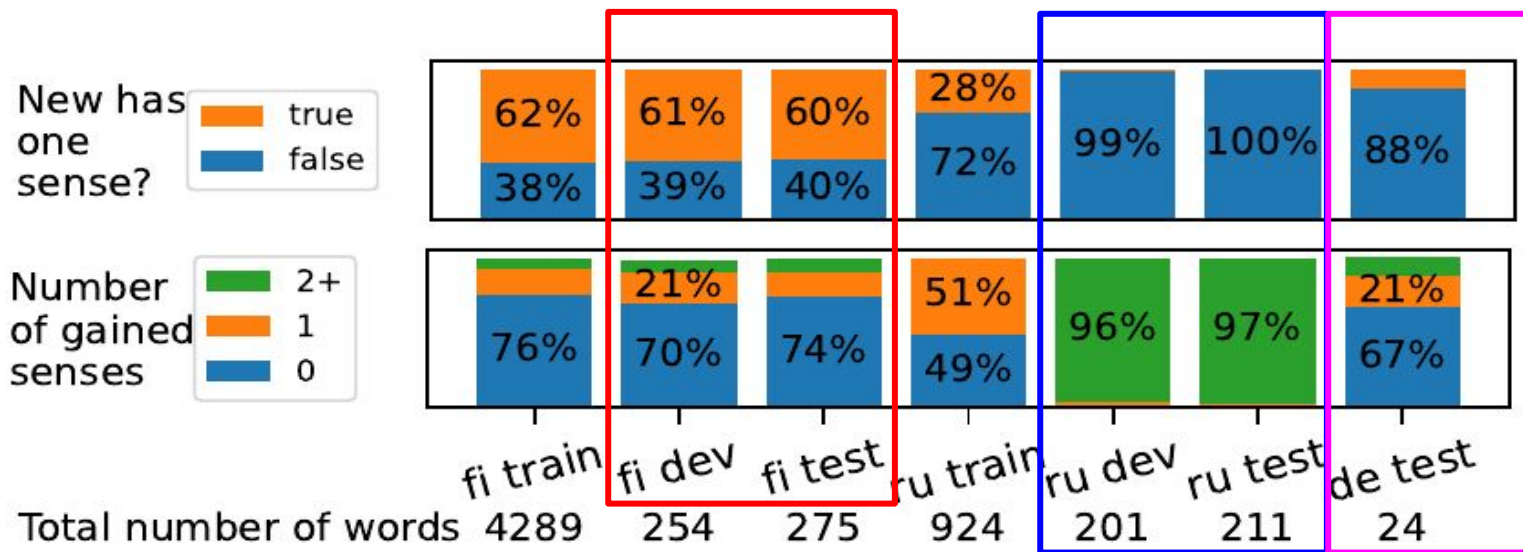
◇ - within 5% (not percentage points) from the best

In Ru words almost always have 2 and more gained senses, so annotating with old senses is not enough – need WSI,
 but in Fi ~70% words do not have gained senses (useless to do anything beyond WSD) and 60% have all new usages with 1 sense only (WSI usually gives >1 cluster $\Rightarrow$ ARI=0).



In Ru words almost always have 2 and more gained senses, so annotating with old senses is not enough – need WSI,
but in Fi ~70% words do not have gained senses (useless to do anything beyond WSD) and 60% have all new usages with 1 sense only (WSI usually gives >1 cluster $\Rightarrow$ ARI=0).

In De 67% of words also do not have gained senses, but only 12% have 1 sense only $\Rightarrow$ the difference in ARI is smaller than for Fi, but still large (maybe when using WSD instead of WSI sense definitions help for better grouping?)



ARI: how well new usages are grouped by their meaning?

WSI gives much better grouping than WSD for Ru, but much worse for Fi and a bit worse for De

Outlier2Cluster preserves the best ARI for Fi and almost best for Ru; a small drop for De (with the Finnish NSD)

Method	Fi	Ru	De	FiRu	AVG
WSD methods					
GR	0.581	0.041	0.386	0.311	0.336
GR FiEnRu	◇0.649	0.048	◇0.521	0.348	0.406
GR Ru	0.568	0.053	0.464	0.310	0.361
GR Fi SG	◇0.638	0.059	◇0.543	0.348	0.413
WSI methods					
Agglomerative	0.209	◇0.259	0.316	0.234	0.261
SCM methods					
AggloM	0.581	0	0.492	0.290	0.357
AggloM FiEnRu	◇0.631	0	0.485	0.315	0.372
Cluster2Sense	0.209	◇0.259	0.316	0.234	0.261
Outlier2Cluster $^{ru}_{fi}$	◇0.649	◇0.247	$\begin{matrix} 0.322 \\ 0.480 \end{matrix}$	◇0.448	$\begin{matrix} 0.406 \\ \text{◇0.459} \end{matrix}$

Underlined - the best of all

Bold - the best in group

◇ - within 5% (not percentage points) from the best

ARI: how well new usages are grouped by their meaning?

AggloM significantly improves upon Agglomerative, works surprisingly well given its simplicity (but only when old usages are available)

Cluster2Sense inherits clustering and its ARI from Agglomerative

Method	ARI				
	Fi	Ru	De	FiRu	AVG
WSD methods					
GR	0.581	0.041	0.386	0.311	0.336
GR FiEnRu	◇ <u>0.649</u>	0.048	◇0.521	0.348	0.406
GR Ru	0.568	0.053	0.464	0.310	0.361
GR Fi SG	◇0.638	0.059	◇ <u>0.543</u>	0.348	0.413
WSI methods					
Agglomerative	0.209	◇ <u>0.259</u>	0.316	0.234	0.261
SCM methods					
AggloM	0.581	0	0.492	0.290	0.357
AggloM FiEnRu	◇0.631	0	0.485	0.315	0.372
Cluster2Sense	0.209	◇ <u>0.259</u>	0.316	0.234	0.261
Outlier2Cluster _{fi} ^{ru}	◇ <u>0.649</u>	◇0.247	^{0.322} 0.480	◇ <u>0.448</u>	^{0.406} ◇ <u>0.459</u>

Underlined - the best of all

Bold - the best in group

◇ - within 5% (not percentage points) from the best

F1: how well usages of old senses are labeled with their senses?

GR FiEnRu – best WSD for usages of old senses for Fi and Ru, almost best for De

Outlier2Cluster preserves F1 for Fi and De (almost no positive predictions), F1 for Ru becomes significantly worse, but still comparable to the best result of other teams

Method	Fi	Ru	F1 De	FiRu	AVG
WSD methods					
GR	0.690	◊0.721	0.694	0.706	0.702
GR FiEnRu	◊ 0.756	◊ 0.750	◊0.745	◊ 0.753	◊ 0.750
GR Ru	0.568	◊ 0.750	◊0.724	0.659	0.681
GR Fi SG	◊0.752	◊0.729	◊ 0.758	◊0.741	◊0.746
WSI methods					
Agglomerative	0.055	0.152	0.042	0.104	0.083
SCM methods					
AggloM	0.674	0	0.695	0.337	0.456
AggloM FiEnRu	◊0.731	0	0.639	0.366	0.457
Cluster2Sense	0.392	0.346	0.432	0.369	0.390
Outlier2Cluster ^{ru} _{fi}	◊ 0.756	0.645	^{0.510} ◊ 0.745	0.701	^{0.637} ◊ 0.715
Other teams					
Holotniekat	0.655	0.661	0.608	0.658	0.641

F1: how well usages of old senses are labeled with their senses?

GR FiEnRu – best WSD for usages of old senses for Fi and Ru, almost best for De

Outlier2Cluster preserves F1 for Fi and De (almost no positive predictions), F1 for Ru becomes significantly worse, but still comparable to the best result of other teams

AggloM shows a bit worse results for Fi and De, comparable to the best of other teams, but does not work for Ru

Cluster2Sense falls behind: bad relabeling or bad grouping inherited from WSI?

Method	Fi	Ru	F1 De	FiRu	AVG
WSD methods					
GR	0.690	◊0.721	0.694	0.706	0.702
GR FiEnRu	◊ 0.756	◊ 0.750	◊0.745	◊ 0.753	◊ 0.750
GR Ru	0.568	◊ 0.750	◊0.724	0.659	0.681
GR Fi SG	◊0.752	◊0.729	◊ 0.758	◊0.741	◊0.746
WSI methods					
Agglomerative	0.055	0.152	0.042	0.104	0.083
SCM methods					
AggloM	0.674	0	0.695	0.337	0.456
AggloM FiEnRu	◊0.731	0	0.639	0.366	0.457
Cluster2Sense	0.392	0.346	0.432	0.369	0.390
Outlier2Cluster ^{ru} _{fi}	◊ 0.756	0.645	^{0.510} ◊0.745	0.701	^{0.637} ◊0.715
Other teams					
Holotniekat	0.655	0.661	0.608	0.658	0.641

Why SCM methods get worse F1 (when not falling back to WSD)?

- Intuitively, F1 of old senses should improve when we try to clean old senses from the usages of obtained senses, but
1. it is calculated for usages of old senses only, doesn't care if usages of gained senses are incorrectly put there (for most words) $\Rightarrow$ only need good WSD
 2. Large penalty if a single usage of some old sense is attributed to a gained sense* $\Rightarrow$ trying to return anything except for the old sense labels hurts very much unless done ideally (but even then it should not help – see 1)

* Assume the target word has k old senses. In case when only old senses are predicted: $F = \frac{F_1 + \dots + F_k}{k}$. If we replace one of the correct predictions of sense 1 with an incorrect prediction of a gained sense: $F' = \frac{F'_1 + \dots + F_k + 0}{k+1} < \frac{F_1 + \dots + F_k + 0}{k+1}$. The drop in this metric is $\frac{F}{F'} > \frac{k+1}{k}$. E.g. in the case $k = 1$, which is a frequent case in the Finnish AXOLOTL-24 dataset, an incorrect prediction of a gained sense for a single usage results in more than 2x decrease of the F1 score.

```
test_usages_predicted_senses = [  
    "novel" if e1 not in old_senses else e1  
    for e1 in test_usages_predicted_senses  
]  
f1 = f1_score(  
    test_usages_gold_senses,  
    test_usages_predicted_senses,  
    average="macro",  
    zero_division=0.0,  
)
```

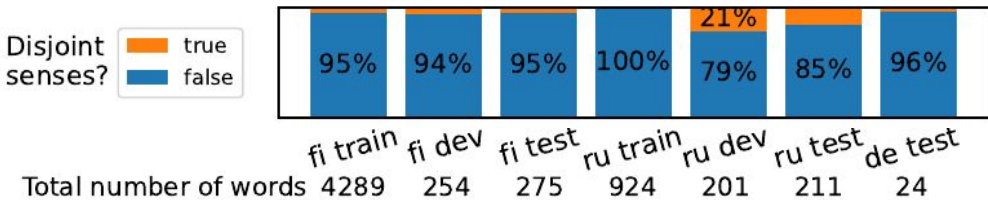
Official eval script:

https://github.com/ltgoslo/axolotl24_shared_task/blob/main/code/evaluation/scorer_track1.py#L95

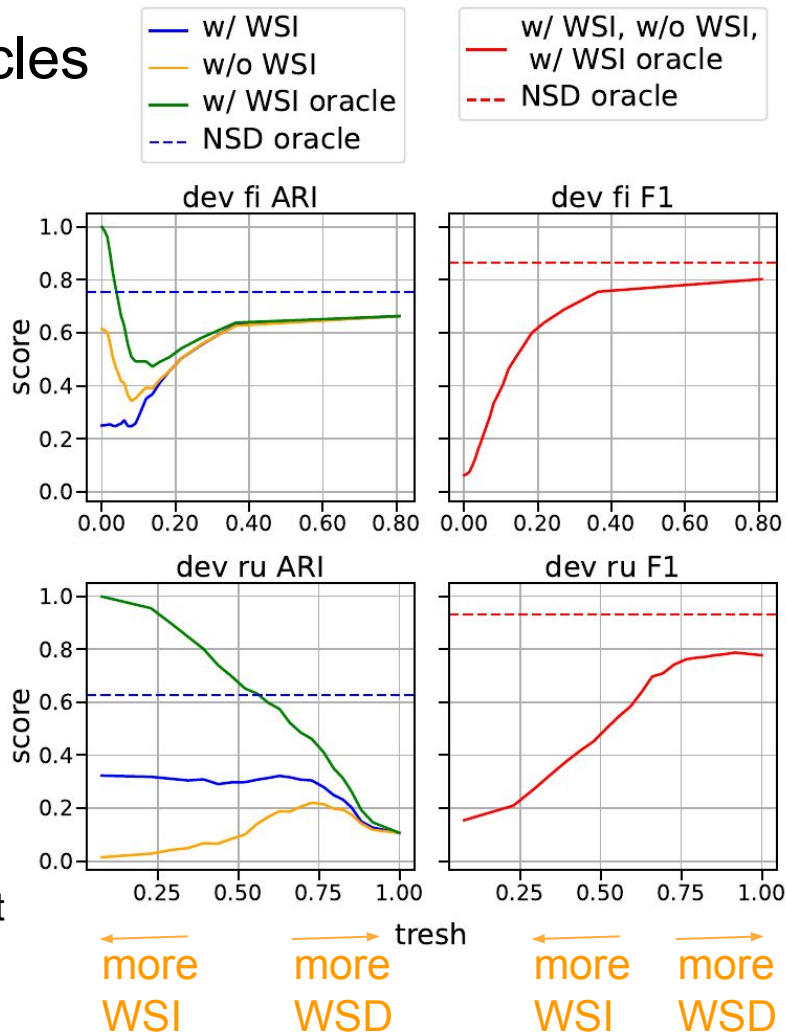
Outlier2Cluster: the threshold and oracles

Trying to clean old senses from usages of gained senses hurts F1, and also ARI on Finnish. For Russian ARI it helps!

We said even ideal NSD shouldn't help for F1, but NSD oracle improves F1?! The effect of words with no new usages of old senses: arbitrarily F1=1 when **all** usages are recognized as usages of gained senses, 0 otherwise!



WSI oracle improves ARI for Russian (>1 gained sense for 97% of words), but not Finnish (only for 7% of words). Similarly, putting all outliers to 1 cluster hurts for Russian, but not Finnish.



GR fine-tuning on AXOLOTL training data

For Fi fine-tuning GR on Fi data is very important (when using as WSD system or for contextulized embeddings in AggloM)! Best when fine-tuned on 3 train sets.

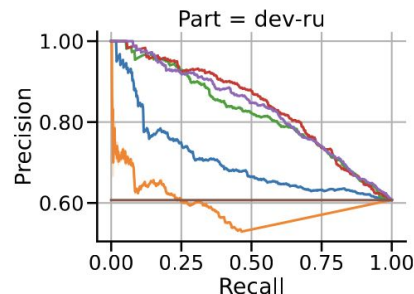
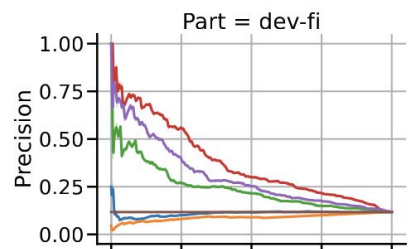
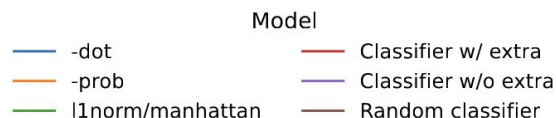
For Ru fine-tuning on Ru data helps a little bit.

For De fine-tuning helps independently of the train set, but mostly on Fi with SG ?!

Method	Fi	Ru	ARI De	FiRu	AVG	Fi	Ru	F1 De	FiRu	AVG
WSD methods										
GR	0.581	0.041	0.386	0.311	0.336	0.690	0.721	0.694	0.706	0.702
GR FiEnRu	0.649	0.048	0.521	0.348	0.406	0.756	0.750	0.745	0.753	0.750
GR Ru	0.568	0.053	0.464	0.310	0.361	0.568	0.750	0.724	0.659	0.681
GR Fi SG	0.638	0.059	0.543	0.348	0.413	0.752	0.729	0.758	0.741	0.746
WSI methods										
Agglomerative	0.209	0.259	0.316	0.234	0.261	0.055	0.152	0.042	0.104	0.083
SCM methods										
AggloM	0.581	0	0.492	0.290	0.357	0.674	0	0.695	0.337	0.456
AggloM FiEnRu	0.631	0	0.485	0.315	0.372	0.731	0	0.639	0.366	0.457
Cluster2Sense	0.209	0.259	0.316	0.234	0.261	0.392	0.346	0.432	0.369	0.390
Outlier2Cluster _{ru} _{fi}	0.649	0.247	0.322 0.480	0.448	0.406 0.459	0.756	0.645	0.510 0.745	0.701	0.637 0.715

5. Experiments with NSD

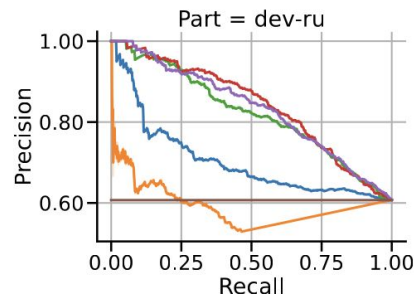
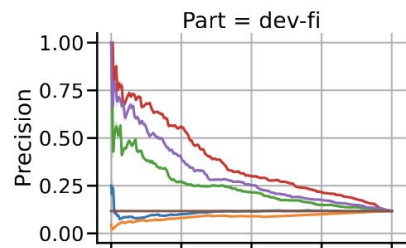
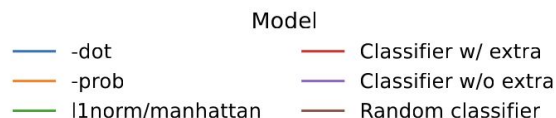
NSD model ablation study



Model	dev fi AP		dev ru AP	
	GR	GR FiEnRu	GR	GR FiEnRu
single features				
cosine	0.106	0.110	0.685	0.695
euclid.	0.106	0.110	0.684	0.694
l2/euclid.	0.106	0.110	0.685	0.695
manh.	0.106	0.113	0.685	0.690
l1/manh.	0.154	0.242	0.816	0.822
full classifiers				
classifier w/ extra	0.378		0.840	
classifier w/o extra	0.305		0.833	
best pairs of features w/o extra features				
l1/manh. + euclid.	0.194	0.284	0.818	0.823
l1/manh. + l2/euclid.	0.195	0.284	0.818	0.823
l1/manh. + manh.	0.192	0.277	0.819	0.823
best pairs of features w/ extra features				
l1/manh. + #old usages	0.190	0.291	0.820	0.827
l1/manh. + #new usages	0.153	0.249	0.821	0.829
#new usages + #old senses	0.266	0.266	0.643	0.643

Underlined - the best of all
Bold - the best in group

NSD model ablation study. Best distance

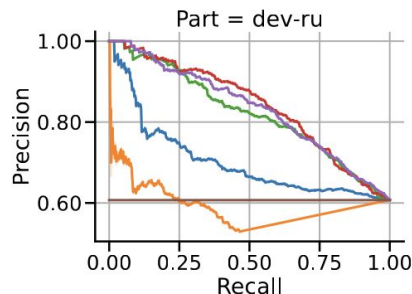
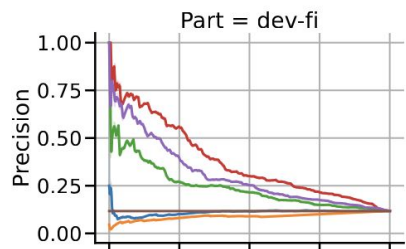
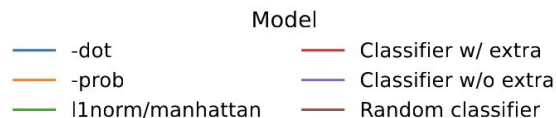


Model	dev fi AP		dev ru AP	
	GR	GR FiEnRu	GR	GR FiEnRu
single features				
cosine	0.106	0.110	0.685	0.695
euclid.	0.106	0.110	0.684	0.694
l2/euclid.	0.106	0.110	0.685	0.695
manh.	0.106	0.113	0.685	0.690
l1/manh.	0.154	0.242	0.816	0.822
full classifiers				
classifier w/ extra	0.378		0.840	
classifier w/o extra	0.305		0.833	
best pairs of features w/o extra features				
l1/manh. + euclid.	0.194	0.284	0.818	0.823
l1/manh. + l2/euclid.	0.195	0.284	0.818	0.823
l1/manh. + manh.	0.192	0.277	0.819	0.823
best pairs of features w/ extra features				
l1/manh. + #old usages	0.190	0.291	0.820	0.827
l1/manh. + #new usages	0.153	0.249	0.821	0.829
#new usages + #old senses	0.266	0.266	0.643	0.643

Underlined - the best of all
Bold - the best in group

the predicted probability of the selected gloss and dot product perform much worse than a classifier. On Finnish they are comparable to a random classifier.

NSD model ablation study. Best distance



Model	dev fi AP		dev ru AP	
	GR	GR FiEnRu	GR	GR FiEnRu
single features				
cosine	0.106	0.110	0.685	0.695
euclid.	0.106	0.110	0.684	0.694
l2/euclid.	0.106	0.110	0.685	0.695
manh.	0.106	0.113	0.685	0.690
l1/manh.	0.154	0.242	0.816	0.822
full classifiers				
classifier w/ extra		0.378		0.840
classifier w/o extra		0.305		0.833
best pairs of features w/o extra features				
<u>l1/manh.</u> + euclid.	0.194	0.284	0.818	0.823
<u>l1/manh.</u> + l2/euclid.	0.195	0.284	0.818	0.823
<u>l1/manh.</u> + manh.	0.192	0.277	0.819	0.823
best pairs of features w/ extra features				
<u>l1/manh.</u> + #old usages	0.190	0.291	0.820	0.827
<u>l1/manh.</u> + #new usages	0.153	0.249	0.821	0.829
#new usages + #old senses	0.266	0.266	0.643	0.643

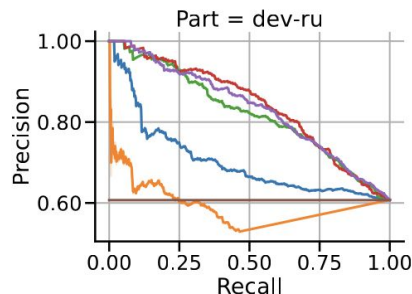
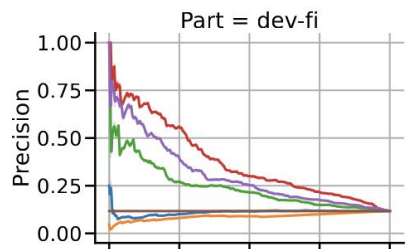
l1/manh.
performs way
better than
other distances

Underlined - the best of all
Bold - the best in group

NSD model ablation study. GR finetuning on AXOLOTL data

Model

- dot
- prob
- l1norm/manhattan
- Classifier w/ extra
- Classifier w/o extra
- Random classifier

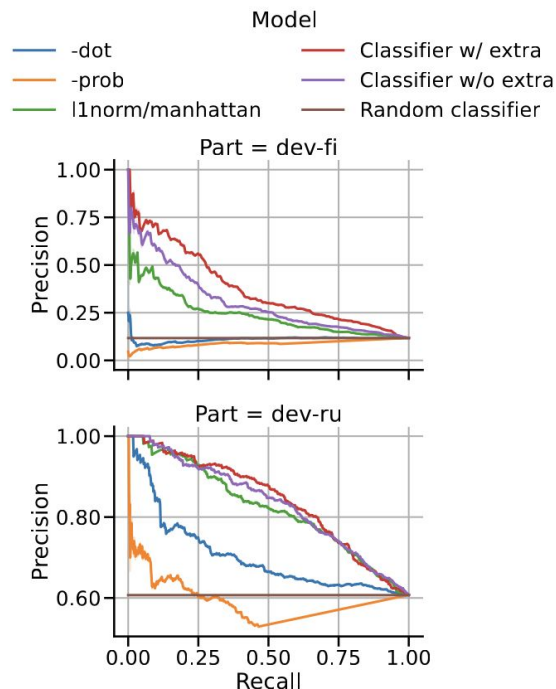


Model	dev fi AP		dev ru AP	
	GR	GR FiEnRu	GR	GR FiEnRu
single features				
cosine	0.106	0.110	0.685	0.695
euclid.	0.106	0.110	0.684	0.694
l2/euclid.	0.106	0.110	0.685	0.695
manh.	0.106	0.113	0.685	0.690
l1/manh.	0.154	0.242	0.816	0.822
full classifiers				
classifier w/ extra	0.378		0.840	
classifier w/o extra	0.305		0.833	
best pairs of features w/o extra features				
l1/manh. + euclid.	0.194	0.284	0.818	0.823
l1/manh. + l2/euclid.	0.195	0.284	0.818	0.823
l1/manh. + manh.	0.192	0.277	0.819	0.823
best pairs of features w/ extra features				
l1/manh. + #old usages	0.190	0.291	0.820	0.827
l1/manh. + #new usages	0.153	0.249	0.821	0.829
#new usages + #old senses	0.266	0.266	0.643	0.643

for the Finish dataset
GlossReader provides
poor embedding
without fine-tuning

Underlined - the best of all
Bold - the best in group

NSD model ablation study. Extra features



Model	dev fi AP		dev ru AP	
	GR	GR FiEnRu	GR	GR FiEnRu
single features				
cosine	0.106	0.110	0.685	0.695
euclid.	0.106	0.110	0.684	0.694
l2/euclid.	0.106	0.110	0.685	0.695
manh.	0.106	0.113	0.685	0.690
l1/manh.	0.154	0.242	0.816	0.822
full classifiers				
classifier w/ extra	0.378		0.840	
classifier w/o extra	0.305		0.833	
best pairs of features w/o extra features				
l1/manh. + euclid.	0.194	0.284	0.818	0.823
l1/manh. + l2/euclid.	0.195	0.284	0.818	0.823
l1/manh. + manh.	0.192	0.277	0.819	0.823
best pairs of features w/ extra features				
l1/manh. + #old usages	0.190	0.291	0.820	0.827
l1/manh. + #new usages	0.153	0.249	0.821	0.829
#new usages + #old senses	0.266	0.266	0.643	0.643

extra features improve NSD model, especially on Finnish

Underlined - the best of all
Bold - the best in group

Conclusions: models

1. Pure WSD give the best F1 scores, but bad ARI for Ru
2. Outlier2Cluster is a reasonable full solution of Subtask 1:
 - * for Ru almost ARI of WSI with a moderate drop in F1-score
 - * for Fi, De selected WSD predictions for almost all usages $\Rightarrow$ almost no drop

$\Rightarrow$ SOTA average ARI, worse average F1 than pure WSD gives but still better than other participants.
3. We have to give WSI the ability to return 1 cluster!
4. NSD is too simple now, probably this is the bottleneck.
5. Combining definitions and example usages can lead to better representations of old senses.

Conclusions: datasets and metrics

1. F1-score (the way it is calculated in the task) promotes pure WSD systems, trying to detect usages of gained senses and not assign old sense to them hurts F1-score (unless done perfectly). But this can be fixed :)

2. To achieve perfect ARI, WSD is enough for words with

- 1 new and 1 old sense $\Rightarrow$ WSD will give 1 cluster, $ARI=1$ ($\Rightarrow$ high ARI on Fi)
- words with no gained senses.

These words constitute 78% of Fi test, 71% of De test.

Only ARI on Ru requires going beyond pure WSD.

	test fi	test ru	test de
words with 1 new and 1 old sense	0.502	0.005	0.042
words with no gained senses	0.742	0.000	0.667
either	0.782	0.005	0.708