APODICTUS

Felix Blessing[1] Julian Kaufmann[2] Johannes Sax[3]

[1]Automatic Prioritization of Dictionary
Update Candidates

[2]Usage Retrieval for Dictionary Headwords
with Applications in Unknown Sense Detection

[3]Sense Definition Generation
and how it can improve WSD

October 14, 2025

Project Introduction

- **Motivation/Background**
  - Language constantly changes
  - ⇒ Need to identify new senses and update dictionary
  - Oxford English Dictionary maintains internal database LEMUR with sense proposals
  - Editors score sense proposals manually
- **Aim of our project**
  - Automate scoring process

Project Overview

3 Main Parts:

1. Usage Retrieval from the NOW corpus
2. Find evidence of sense proposals in usages and assign prioritization scores
3. Sense Definition Generation for unrecorded senses

Automatic Prioritization Of DICTionary Update candidateS

Task

- **Input**
  - LEMUR database L containing sense proposals $s_p \in L$
  - Set of Usages U of sense proposal target words
  - Dictionary D containing senses $s \in D$
- **Output**
  - Prioritization scores $p(s_p)$ for each sense proposal s_p, based on evidence found in U

Data: Dictionaries

- 1300 LEMUR sense proposals

sense_id	lemma	gloss
LMR2-81764	spam	Slang. To press or strike (a computer key, button, etc.) many times in quick succession.

Table: LEMUR sense proposal for "spam"

- ODE dictionary entries associated with LEMUR sense proposals

sense_id	lemma	gloss
spam_006	spam	irrelevant or unsolicited messages sent over the internet, typically to a large number of users, for the purposes of advertising, phishing, spreading malware, etc.
spam_009	spam	a tinned meat product made mainly from ham
spam_013	spam	send the same message indiscriminately to (a large number of internet users)

Table: ODE Dictionary entries for "spam"

Data: Usages

- Usages of sense proposal words

identifier	lemma	usage
NOW-17060	spam	In dramatic sequences, God of War might ask the player to spam "X" or twirl the control sticks to mimic the action happening on screen
NOW-18010	spam	Spam, trout, fried chicken, moon pies and anything slathered in mayonnaise – those are some of the flavors of South Korea's home cooking that might seem just a bit familiar to the U.S.
NOW-17061	spam	For big, elaborate boss battles, Barlog said, players can expect the "Track and Field" design, referring to the classic NES game in which players quickly spammed buttons to create a feeling of physical exertion

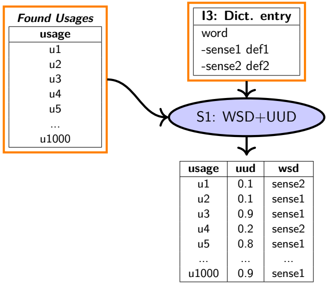
Table: Example usages for target word "spam".

Step 1: Filter Recorded Usages

- Filter usages containing already recorded dictionary senses

$\Rightarrow$ Compare usages with main dictionary

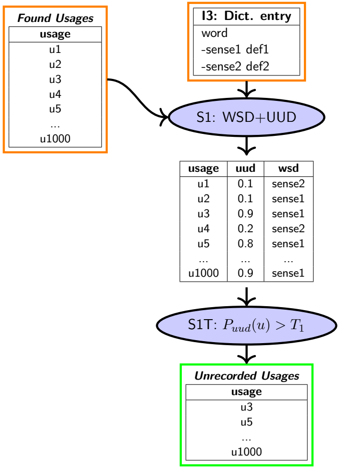
Step 1: Filter Recorded Usages



- Filter usages containing already recorded dictionary senses

$\Rightarrow$ Compare usages with main dictionary

Step 1: Filter Recorded Usages



- Filter usages containing already recorded dictionary senses

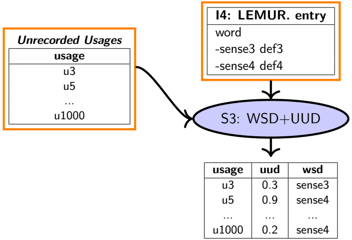
$\Rightarrow$ Compare usages with main dictionary

Step 3: Find LEMUR Evidence

- Search for LEMUR senses in remaining unrecorded usages

$\Rightarrow$ Compare Usages with LEMUR sense proposals

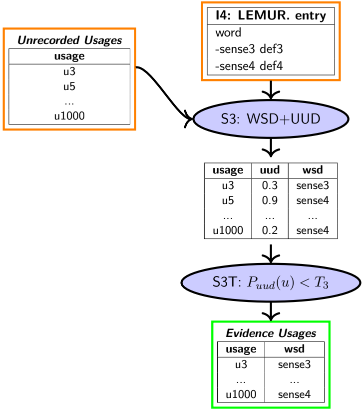
Step 3: Find LEMUR Evidence



- Search for LEMUR senses in remaining unrecorded usages

$\Rightarrow$ Compare Usages with LEMUR sense proposals

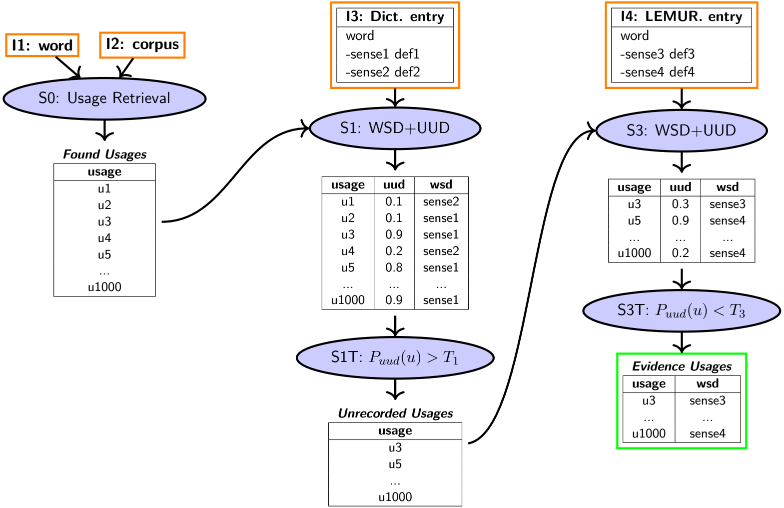
Step 3: Find LEMUR Evidence



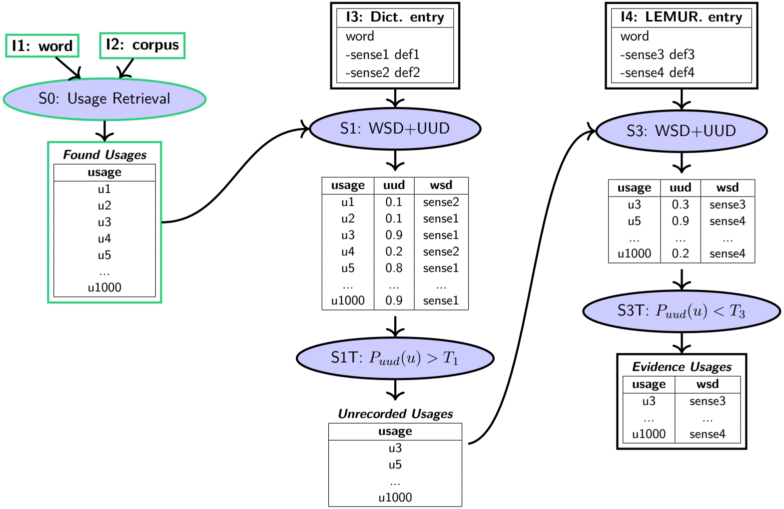
- Search for LEMUR senses in remaining unrecorded usages

$\Rightarrow$ Compare Usages with LEMUR sense proposals

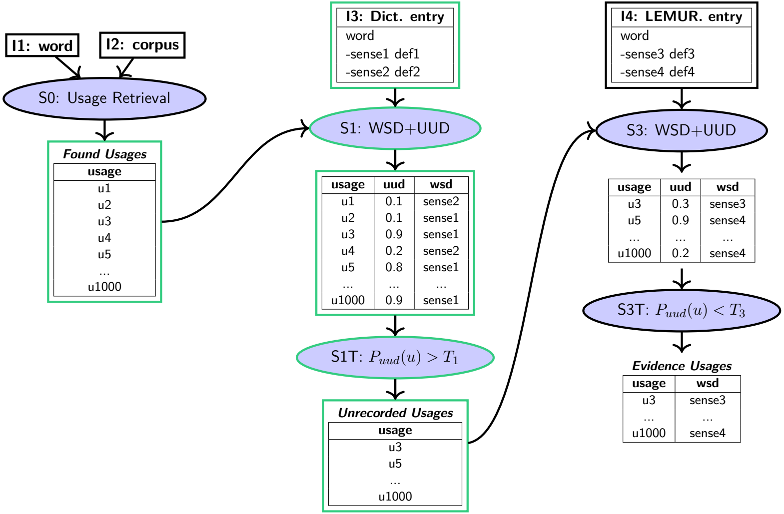
Pipeline Overview



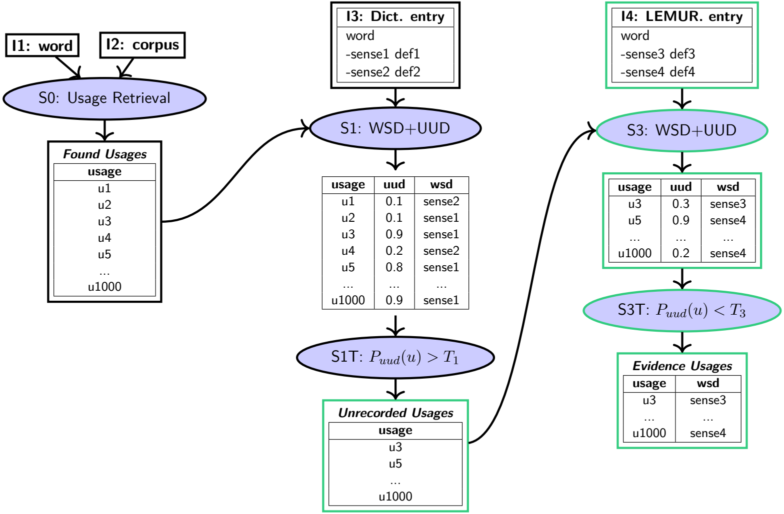
S0: Usage Extraction



S1: Filter Recorded Usages



S3: Find LEMUR evidence



Model: Output

lemma	sense_id	total_usages	evidence_count	evidence_ratio	gloss	source
spam	LMR2-81764	7244	15	0.0021	...	LEMUR100

Table: evidence.tsv file containing results per sense

- **total_usages** = Total number of given usages for the target word
- **evidence_count** = Number of usages assigned to this LEMUR sense proposal
- **evidence_ratio** = $\frac{\text{evidence\_count}}{\text{total\_usages}}$

Outlier2Cluster

- Method proposed by Kokosinskii et al.[1]
- Originally designed for Shared Task involving Semantic Change Detection [2]
- Adapted to our task using a wrapper
- **How it works:**
  - Creates embedding vectors for glosses and usages
  - WSD: Assign to each usage the most suitable sense (dot product)
  - UUD: Given the usage and the most suitable sense calculate outlier probability (logistic regression function)
  - Apply threshold

Outlier2Cluster

Logistic Regression Classifier weights : **own_weights**

- Trained on 100 annotated usages
- 2 words, 50 usages each

Full Pipeline Run

Parameter	Value
Sense Proposals	All 1300 LEMUR sense proposals
Threshold S1	0.19
Threshold S3	0.4
Max usages per word	10,000
NSD classifier	`own_weights`

Quality Control Annotation

- Sample 15 in-ODE and 15 out-of-ODE words with at least 1 LEMUR evidence
- For each word sample up to 10 LEMUR prediction usages per probability-bin $[0 - 0.1], [0.1 - 0.2], [0.2 - 0.3], [0.3 - 0.4]$

Quality Control Annotation

- Sample 15 in-ODE and 15 out-of-ODE words with at least 1 LEMUR evidence
- For each word sample up to 10 LEMUR prediction usages per probability-bin $[0-0.1], [0.1-0.2], [0.2-0.3], [0.3-0.4]$

sense_id	**lemma**	**gloss**
LMR2-81764	spam	Slang. To press or strike (a computer key, button, etc.) many times in quick succession.

Table: LEMUR sense proposal for “spam”

label	**usage**
	might ask the player to **spam** "X" or twirl the control sticks
	players quickly **spammed** buttons
	click the "X" to report **spam** or abuse.

Table: Sampled “spam” usages

Quality Control Annotation

- Sample 15 in-ODE and 15 out-of-ODE words with at least 1 LEMUR evidence
- For each word sample up to 10 LEMUR prediction usages per probability-bin $[0-0.1], [0.1-0.2], [0.2-0.3], [0.3-0.4]$

sense_id	lemma	gloss
LMR2-81764	spam	Slang. To press or strike (a computer key, button, etc.) many times in quick succession.

Table: LEMUR sense proposal for “spam”

label	usage
1	might ask the player to **spam** "X" or twirl the control sticks players quickly **spammed** buttons click the "X" to report **spam** or abuse.

Table: Sampled “spam” usages

Quality Control Annotation

- Sample 15 in-ODE and 15 out-of-ODE words with at least 1 LEMUR evidence
- For each word sample up to 10 LEMUR prediction usages per probability-bin $[0 - 0.1], [0.1 - 0.2], [0.2 - 0.3], [0.3 - 0.4]$

sense_id	**lemma**	**gloss**
LMR2-81764	spam	Slang. To press or strike (a computer key, button, etc.) many times in quick succession.

Table: LEMUR sense proposal for “spam”

label	**usage**
1	might ask the player to **spam** "X" or twirl the control sticks
1	players quickly **spammed** buttons
	click the "X" to report **spam** or abuse.

Table: Sampled “spam” usages

Quality Control Annotation

- Sample 15 in-ODE and 15 out-of-ODE words with at least 1 LEMUR evidence
- For each word sample up to 10 LEMUR prediction usages per probability-bin $[0-0.1], [0.1-0.2], [0.2-0.3], [0.3-0.4]$

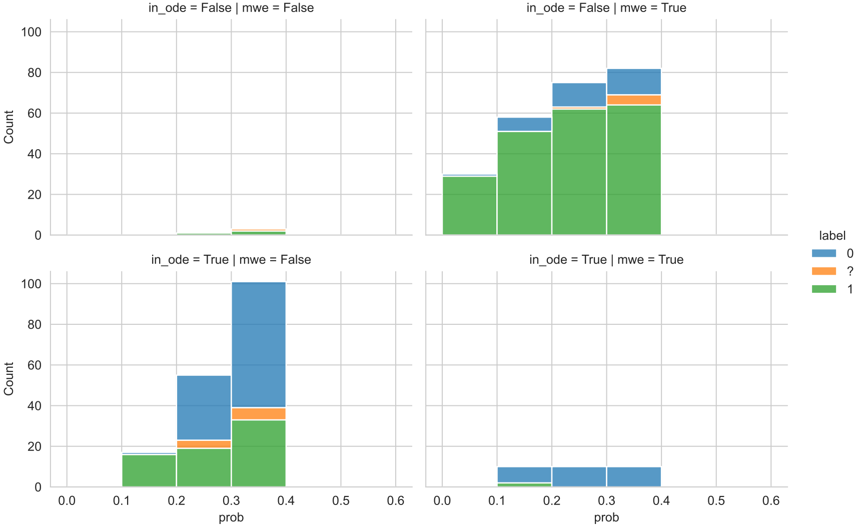
sense_id	**lemma**	**gloss**
LMR2-81764	spam	Slang. To press or strike (a computer key, button, etc.) many times in quick succession.

Table: LEMUR sense proposal for “spam”

label	**usage**
1	might ask the player to **spam** "X" or twirl the control sticks
1	players quickly **spammed** buttons
0	click the "X" to report **spam** or abuse.

Table: Sampled “spam” usages

Quality Control Annotation



Error Analysis

Impact of Error Types

Error Type	Affected Usages	Affected Senses
All Errors (False Positives)	156	19
Loose Lexical or Semantic Overlap	61 (39.1%)	15
POS Mismatch	19 (12.2%)	6
Corpus Artifacts and Corruption	18 (11.5%)	9
Problematic Definition	–	6

Table: Short LEMUR definition examples

Error Analysis

1. Loose Topical or Lexical Overlap

The Perseids @ @ @ @ @ @ @ @ @ @ behind by the comet Tuttle-Swift on its elongated, 133-year orbit around the Sun. Each meteor is a piece of broken-off comet, which explodes as it hits Earth's atmosphere. Within the broad belt of debris there are also denser dust ribbons created when the comet passes closest to the Sun in its orbit – a juncture called perihelion. This year, Earth is on a collision course with three of the most heavily populated of these trails – created in the years 1862, 1737 and 1479. -' Kamikaze run' - "The meteors you'll see this year are from comet flybys that occurred hundreds if not thousands of years ago," NASA meteoroid expert Bill Cooke said in a statement. "And they've travelled billions of miles before their kamikaze run into Earth's atmosphere." However, there is no risk to our planet. In fact, astronomers' main concern is the weather, with cloud cover predicted for parts of Europe. There @ @ @ @ @ @ @.

Word	**Gloss**
dust ribbon	weather

Table: LEMUR entry

Error Analysis

2. POS mismatch

... Musk carried a sink into Twitter's office. ...

Word	PoS	Gloss
sink into	phrasal verb	intr. To put one's hand into (a pocket)

Table: LEMUR entry

Error Analysis

3. Corpus artifacts and corruption near target word

> … What does the shortage of @ @ @ @ @ @ @ @ @ @ billion promo industry? MV- I think what I am saying is …

> With the approval, Nigeria has 173 universities, out of which 79 of them are private. 2 COMMENTS 2019 Promo Are you into molding, building, and construction this is to inform the general public that individual can now order DangoteCement directly from the factory at a reduce price of …

> … to wipe out malaria in Kenya. ADVERTISEMENT ADVERTISEMENT Currently, the world is largely embroiled in one of the greatest health emergencies …

Error Analysis

4. Problematic LEMUR Definitions

word	definition	problematic characteristic
dust ribbon	weather	short and general
sticker	A person who posts bills, posters, etc.; = STICKER-UP n. \\Cf. 'bill sticker' 'advertisement sticker'	noisy, special characters

Table: Problematic definitions

Development set Dev3

Sample from full pipeline run data:

- Sample 50 In-ODE and 50 Out-of-ODE words
- From extracted usages sample up to 30 for Out-of-ODE words
- From extracted usages sample up to 100 for in-ODE words

Development set Dev3

- Annotate 24 in-ODE and 24 out-of-ODE words
- 2 external annotators, both native english speakers

Case	**Example**
Dictionary sense	`sense_id = 2` or `sense_id = 2,4,3`
New unrecorded sense	`sense_id = -1`
Corrupted usage	`sense_id = x`
Annotator uncertain	`sense_id = 0`

Table: Annotation instructions

Development set Dev3

Metric	Value
Total Usages	2746
In-ODE Usages	2177
Out-of-ODE Usages	569
LEMUR sense Usages	375
LEMUR sense Usages In-ODE	70
LEMUR sense Usages Out-of-ODE	305

Table: Basic analysis of annotations.

Development set Dev3: Annotation Agreement

- Based on 100 common annotated usages.
- Annotator 1,2: main annotators (external)
- Annotator 3: Only for Agreement (internal)

Cohen's κ	**Annotator 2**	**Annotator 3**
Annotator 1	$\kappa_l = 0.978$	$\kappa_l = 0.894$
Annotator 2		$\kappa_l = 0.916$

Krippendorff's α	**Value**
α_l	0.721

- α_l, κ_l: LEMUR usage Y/N

Precision and Recall (In-ODE=False)

LEMUR Sense	lemma	Total LEMUR Senses	Predicted LEMUR Senses	Correct LEMUR Senses	Precision	Recall	In-ODE
LMR2-65777	kanafeh	30	15	15	1.0	0.5	False
LMR2-81261	to thread the needle	14	0	0	-	0.0	False
LMR2-49106	acker	0	0	0	-	-	False
LMR2-61766	air tanker	27	11	11	1.0	0.41	False
LMR2-76433	drinking culture	30	1	1	1.0	0.03	False
LMR2-47292	gold flake	6	0	0	-	0.0	False
LMR2-67070	blanket-like	28	8	8	1.0	0.29	False
LMR2-76273	beer feast	0	0	0	-	-	False
LMR2-56162	capture-the-flag	19	0	0	-	0.0	False
LMR2-74873	chairing	0	0	0	-	-	False
LMR2-60257	Willmore conjecture	1	0	0	-	0.0	False
LMR2-66184	directedness	30	6	6	1.0	0.2	False
LMR2-81027	superheroic	29	0	0	-	0.0	False
LMR2-79454	gravity bong	30	0	0	-	0.0	False
LMR2-696	blue light special	6	0	0	-	0.0	False
LMR2-73446	Occidentalism	18	23	15	0.65	0.83	False
LMR2-15173	speciality rule	2	2	2	1.0	1.0	False
LMR2-33387	Netflix and chill	5	0	0	-	0.0	False
LMR2-50373	dog-hole	0	0	0	-	-	False
LMR2-63695	empanadilla	17	1	1	1.0	0.06	False
LMR2-66010	metrophobia	0	0	0	-	-	False
LMR2-10869	unmixing	5	4	1	0.25	0.2	False
LMR2-35196	sideway	0	0	0	-	-	False
LMR2-70721	fried slice	8	2	1	0.5	0.12	False

Precision and Recall (In-ODE=True)

LEMUR Sense	lemma	Total LEMUR Senses	Predicted LEMUR Senses	Correct LEMUR Senses	Precision	Recall	In-ODE
LMR2-42417	adoptive	7	0	0	-	0.0	True
LMR2-78835	prefill	5	1	0	0.0	0.0	True
LMR2-53661	booby	47	1	1	1.0	0.02	True
LMR2-49027	hale	0	0	0	-	-	True
LMR2-82760	drinker	0	0	0	-	-	True
LMR2-45671	funk	0	0	0	-	-	True
LMR2-64260	buckshee	0	0	0	-	-	True
LMR2-65520	fastball	0	12	0	0.0	-	True
LMR2-48282	ballroom	1	1	0	0.0	0.0	True
LMR2-50622	VOC	2	0	0	-	0.0	True
LMR2-48150	atom	0	0	0	-	-	True
LMR2-64577	beast	1	0	0	-	-	True
LMR2-25261	bump	0	0	0	-	-	True
LMR2-25264	bump	0	0	0	-	-	True
LMR2-11484	large	0	0	0	-	-	True
LMR2-66285	flow	0	0	0	-	-	True
LMR2-65107	hammer	0	0	0	-	-	True
LMR2-44981	Titan	0	0	0	-	-	True
LMR2-13442	versatile	1	0	0	-	-	True
LMR2-75467	craven	4	0	0	-	0.0	True
LMR2-61201	dog biscuit	2	0	0	-	0.0	True
LMR2-58873	annunciate	0	0	0	-	-	True
LMR2-54840	anchor	0	0	0	-	-	True
LMR2-76326	choral	0	0	0	-	-	True

Precision and Recall

Setting	Precision	Recall
Regular	0.7045	0.1653
Regular (In-ODE = Y)	0.0667	0.143
Regular (In-ODE = N)	0.8356	0.2
Macro	0.6716	0.1466
Macro (In-ODE = Y)	0.25	0.003
Macro (In-ODE = N)	0.8402	0.2024

$\Rightarrow$ Promising results but room for improvement

$\Rightarrow$ Out-of-ODE performance: good

$\Rightarrow$ In-ODE performance: unreliable

Hyperparameter analysis

Test different models and thresholds for UUD step.

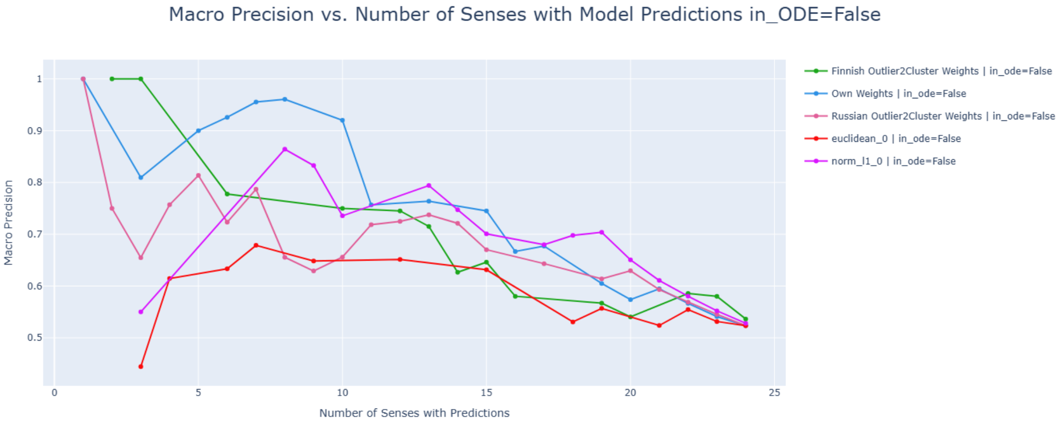
- **Logistic Regression Classifiers**
  - *own_weights*: Own weights trained on 100 annotated usages
  - *Russian Outlier2Cluster Weights*: From the original Outlier2Cluster trained on a Russian development set project [1]
  - *Finnish Outlier2Cluster Weights*: From the original Outlier2Cluster trained on a Finnish development set [1]
- **Single Distance Metrics**
  - Cosine, euclidean, manhattan, L1-Norm (normalized euclidean distance), L2-Norm (normalized manhattan distance)

Hyperparameter analysis

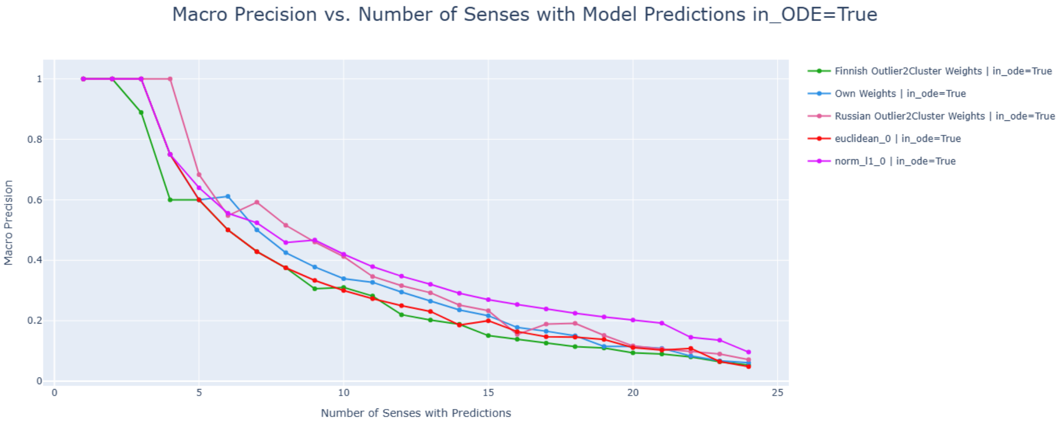
How?

- Grid Search: Test 10.000 S1 and S3 threshold combinations
- Calculate Macro Precision
- Number of words with LEMUR predictions as replacement for recall

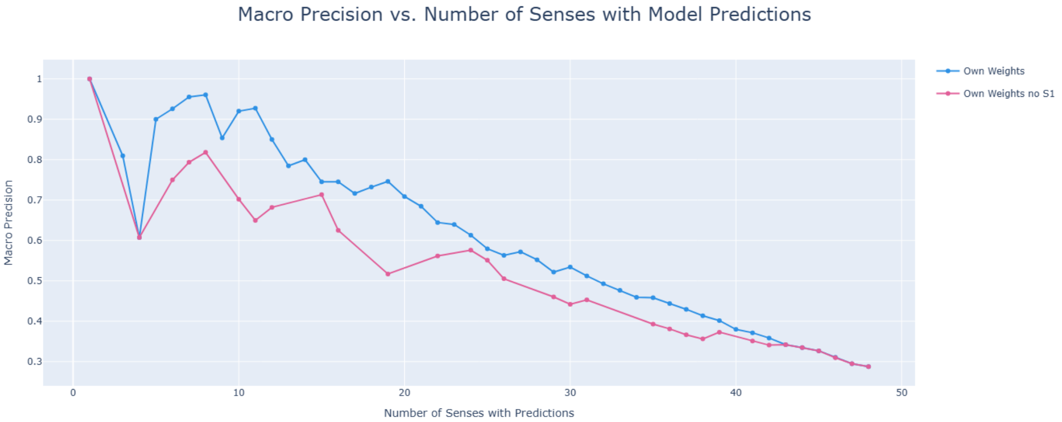
Hyperparameter analysis



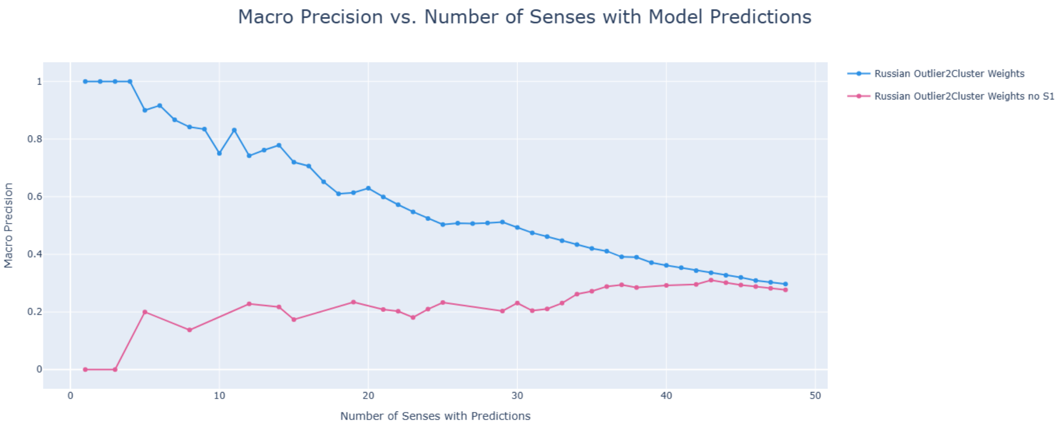
Hyperparameter analysis



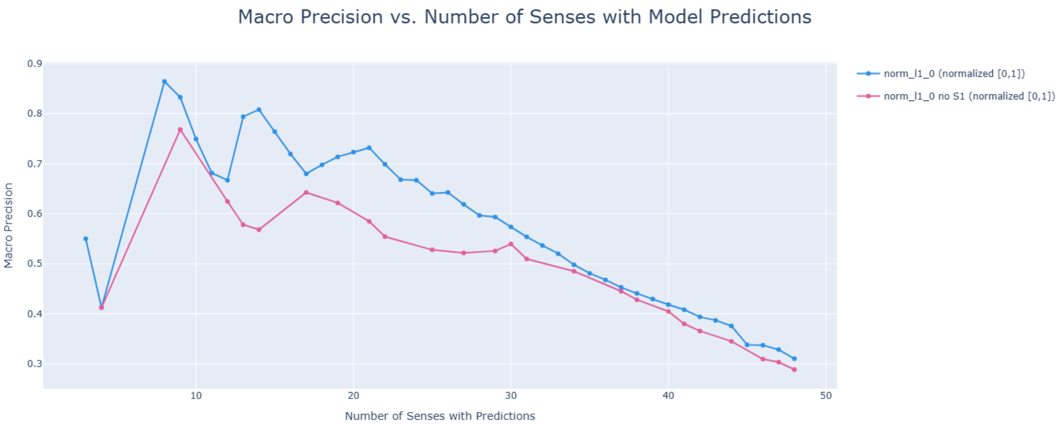
S1: (Filtering) Evaluation



S1: (Filtering) Evaluation

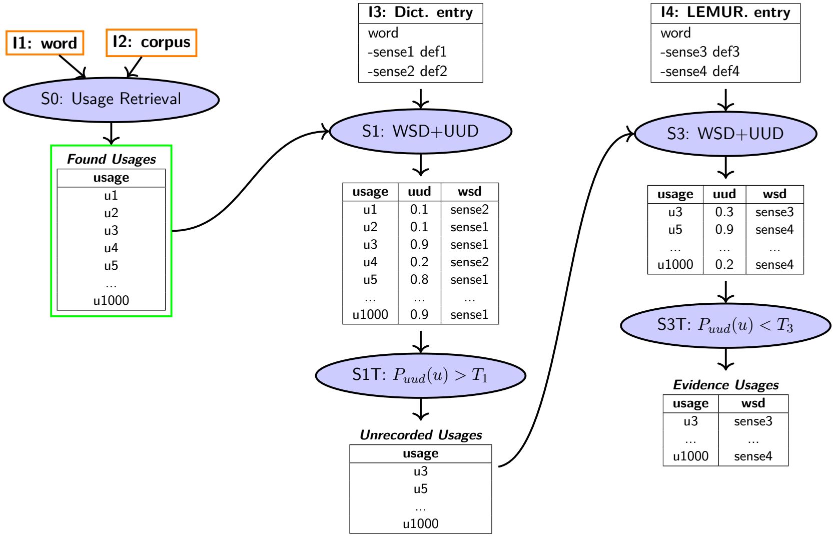


S1: (Filtering) Evaluation

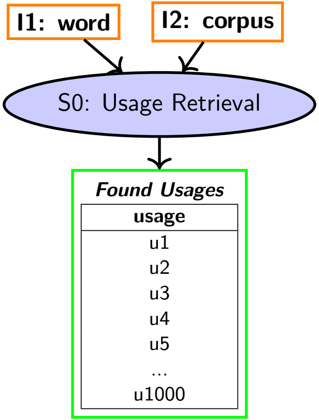


Usage Retrieval

Pipeline Step S0: Usage Retrieval



S0: Overview



Inputs

word: the headword/lemma we are searching for (LEMUR entries)

corpus: the corpus we are searching (NOW corpus)

Output

usages: usages of *word* found in *corpus*

S0: Headword Preprocessing

Some entries contain:

- multiple variants
- abbreviations in brackets
- *the* or *to* suffix
- placeholders like *someone*

LEMUR headword	**Queries**
yerk | *yark*	*yerk* and *yark*
like-as-we/they-lie	*like-as-we-lie* and *like-as-they-lie*
international match point (IMP)	*international match point*
Silent Places, the	*Silent Places*
to come back to haunt someone	*to come back to haunt* #

S0: Corpus

NOW Corpus

> *The NOW corpus (News on the Web) has been created by Mark Davies, and it contains 23.2 billion words of data from web-based newspapers and magazines from 2010 to the present time [...]*
>
> – english-corpora.org

- Texts are scraped from the internet
- Include unwanted artefacts
- Tagged version of the corpus (tokenized and lemmatized)
- Has copyright censoring

S0: Corpus Structure

TextID	TokenID	Word	Lemma	PoS
1334916	262406	@@1334916		fo
1334916	262407	<h>		null
1334916	262408	Britain	britain	np1
1334916	262409	is	be	vbz
1334916	262410	facing	face	vvg
1334916	262411	an	a	at1
1334916	262412	"		"
1334916	262413	obesity	obesity	nn1
1334916	262414	time-bomb	time-bomb	nn1
1334916	262415	"		"

S0: Corpus Structure

Row	Word	Lemma	PoS
1	<h>		null
2	Britain	britain	np1
3	is	be	vbz
4	facing	face	vvg
5	an	a	at1
6	"		"
7	obesity	obesity	nn1
8	time-bomb	time-bomb	nn1
9	"		"

S0: Matching

Row	Word	Lemma	PoS
1	<h>		null
2	Britain	britain	np1
3	is	be	vbz
4	facing	face	vvg
5	an	a	at1
6	"		"
7	obesity	obesity	nn1
8	time-bomb	time-bomb	nn1
9	"		"

S0: Matching

Row	Word	Lemma	PoS
1	<h>		null
2	Britain	britain	np1
3	is	be	vbz
4	facing	face	vvg
5	an	a	at1
6	"		"
7	obesity	obesity	nn1
8	time-bomb	time-bomb	nn1
9	"		"

S0: Outputs

Fragment reassembly

- Join tokens with space
- Exceptions are e.g. punctuation

Examples

`is_VBZ facing_VVG an_AT1`

→ *is␣facing␣an*

`Spam_NN1 ,_y test_VV0`

→ *Spam,␣test* instead of *Spam␣,␣test*

S0: Quotation Marks

Problem

- Original text not available
- Spacing differs at start and end of quote

Solution

→ Mark pairs of quotes

Input	`" " "`
Output	`"start "end "start`

`This isn't "␣easy␣"` NOW-1234GB

`This isn't "easy"` NOW-1234GB

S0: Outputs

Text clean-up

- Remove unwanted artefacts

Input Usage	**Cleaned Usage**
* Spam, spam, and eggs *	*Spam, spam, and eggs*
& < >	*& < >*
*Spam and **123;123;TOOLONG eggs*	*Spam and eggs*
More␣␣and more	*More␣and␣more*

S0: Examples

[...] Quantum computing can help enhance @ @ @ @ @ @ @ @ @ @ variational quantum eigensolver (VQE) algorithm in a quantum simulator to calculate ground state vibrational energies of reactants and products of the CO2 and NH3 reaction. The VQE calculations yield ground vibrational energies of CO2 and NH3 with similar accuracy to classical computing. In the presence of hardware noise, Compact Heuristic for Chemistry (CHC) ***ansatz*** *with shallower circuit depth performs better than Unitary Vibrational Coupled Cluster. The "Zero Noise Extrapolation" error-mitigation approach in combination with CHC ansatz improves the vibrational calculation accuracy. Excited vibrational states are accessed with quantum equation of motion method for CO2 and NH3. [...]*

S0: Examples

[...] Factor XI LICA to Reduce Events Such as Heart Attack and Stroke in Patients Whose Kidneys Are no Longer Able to Work as They Should and Require Treatment to Filter Wastes From the Blood: Focus is on the Safety of BAY2976217 and the Way the Body Absorbs, Distributes and Removes the ***Study Drug*** *(RE-THINc ESRD) Factor XI LICA to Reduce Events Such as Heart Attack and Stroke in Patients Whose Kidneys Are no Longer Able to Work as They Should and Require Treatment to Filter Wastes From the Blood: Focus is on the Safety of BAY2976217 and the Way the Body Absorbs, Distributes and Removes the Study Drug (RE-THINc ESRD) Patients whose kidneys are no longer able to work as they should and require treatment to filter wastes from the blood (hemodialysis) are at high risk for blood clots that form in blood vessels (thrombosis) blocking blood flow that causes heart attacks, strokes, and other life-threatening conditions. [...]*

S0: Deduplication

`there is an update to a comment thread you follow or if a user`	NOW-1234GB
`there is an update to a comment thread you follow or if a user`	NOW-5678US

Identifier	`NOW-1234GB`
Duplicates	2

S0: Incorporating Metadata

- Search text id in corpus metadata
- Add additional information to usages

TextID	1334916
Date	10-01-01
Region	GB
URL	http://www.telegraph.co.uk/news/health/news/6875091/Number-of-people-dying-as-a-result-of-obesity-doubles-in-10-years.html
Title	*Number of people dying as a result of obesity doubles in 10 years*

Table: Metadata for TextID *1334916*

S0: Evaluation

On usages from retrieval run for dev2, including 60 headwords

Recall: Percentage of usages found by retrieval of total usages in corpus

- Median recall of $\approx 94\%$
- Still usages missed by retrieval
- Copyright censoring one factor

		LEMUR		
		300	**1000**	
Type	**SWE**	94.9	100.0	100.0
	MWE	91.8	93.2	91.9
		92.9	100.0	94.2

Table: Median Recall in Percent

S0: Evaluation

Precision: Percentage of correctly matched of total retrieved usages

- Sample up to 5 usages randomly
- Annotated binarily, check if they fit the lemma
- 228/300 usages were sampled
- Precision of 100%

`unforgettable hook and the video is widely shared. Perhaps, with our goldfish memory, we will soon forget about the angry don`	NOW-2311IN-103330153-30817188830

S0: Challenges

Resolved

- Multiple entries per line

  Preprocessing *yerk* / *yark* → *yerk* and *yark*
- "simple" MWE → merge tokens for matching
- MWE with words in-between

  Placeholder *feel someone's pain* → *feel #'s pain*
- Unwanted artefacts

  Text clean-up *<p>Spam</p>* → *Spam*
- Empty lemma column → use lowercased word form

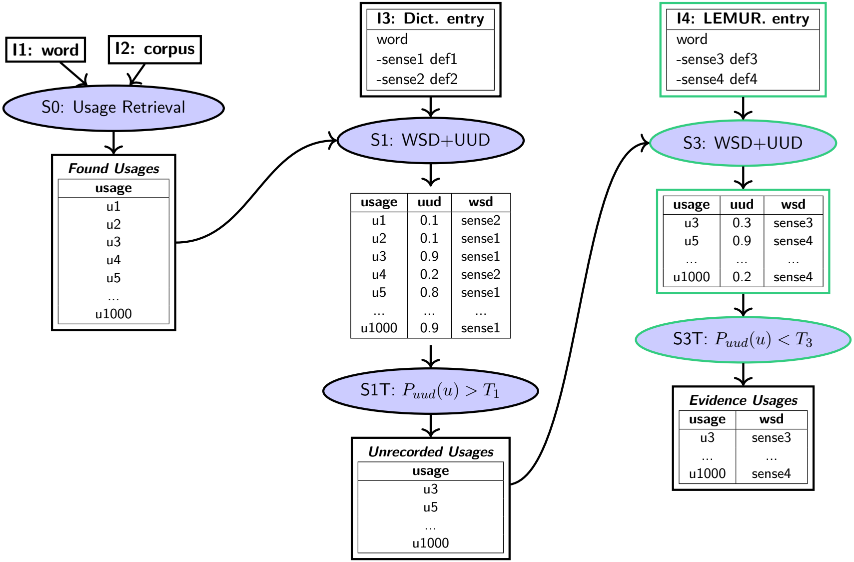
S0: Challenges

Open

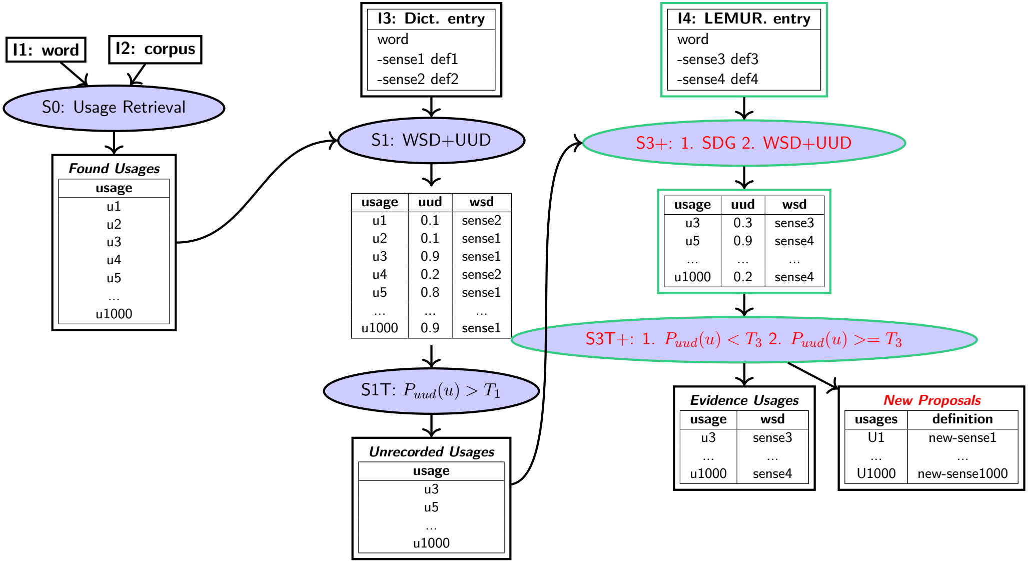
- Spelling variants (from ODE)
- Infrequent PoS
- Headwords with few usages in entire corpus
- Inconsistent quote spacing

Sense Definition Generation

Pipeline SDG: Overview



Pipeline SDG: Overview



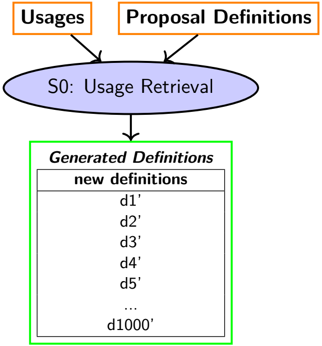
How to use SDG in the pipeline?

1. **Improve definition proposals**
   - proposed definitions often aren't as precise as ODE ones
   - Could improve the quality of WSD+UUD
2. **Create new proposals**
   - Pipeline can detect more unrecorded senses (not just Lemur, ...)
   - Automatic generation of proposals with evidence

Why Sense Definition Generation (SDG) Matters

- **Precise sense definitions**
  - Improve WSD task
  - Human readability
- **Automation:** $\downarrow$ cost, $\uparrow$ speed
  - Manual definition writing is time-consuming and expensive
  - SDG can solve
- **Slang, Medicine, ... :** No one can know everything
  - Slang, regional variation, domain-specific senses, ...
  - SDG can understand and/or knows more

SDG: Task



Inputs

Usages: Usages for headword w

Proposal Definitions: Definition proposals like LEMUR for w

Output

Generated Definitions: New and improved Definitions of *Proposal Definitions*

SDG: Task Description

Input:

- a headword w
- a set of retrieved usages U_w for w
- a (optional) proposed definition d for the new sense s

Output:

- a new/proposed definition d' for the sense s
- d' should accurately reflect the meaning of s

SDG: Example

Dictionary:

Sense ID	**Definition**
cell 1	*Biological cell*
cell 2	*Cell phone*
cell 3	*Prison cell*

Usages:

Context	**Sense ID (gold)**
I'm in a cell.	cell 3
My android cell	cell 2
A onion cell	cell 1

SDG: Generated Definitions

Updated Dictionary:

Sense ID	**Original Definition**	**Generated** $SDG_{model+02c}$
cell 1	*Biological cell*	*The basic structural and functional unit of all organisms.*
cell 2	*Cell phone*	*A portable telephone using radio signals for calls.*
cell 3	*Prison cell*	*A small room used as a place of confinement for prisoners.*

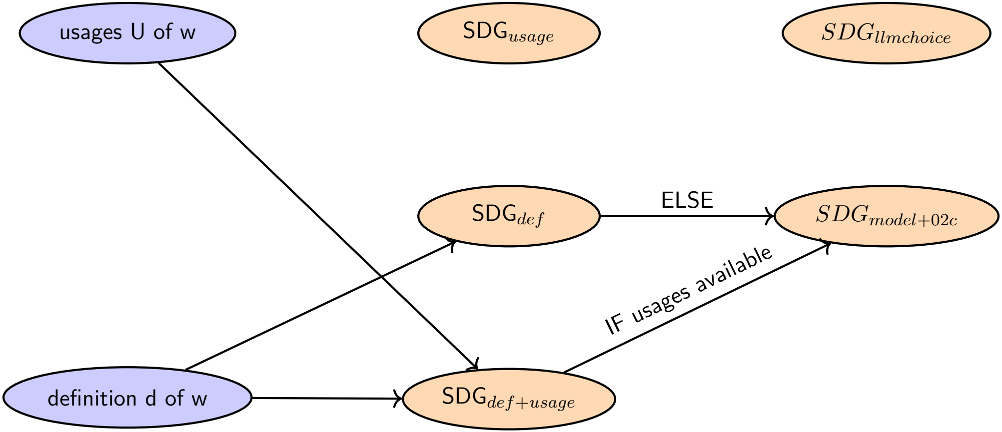
Usages:

Context	**Sense ID (gold)**
I'm in a cell.	cell 3
My android cell	cell 2
A onion cell	cell 1

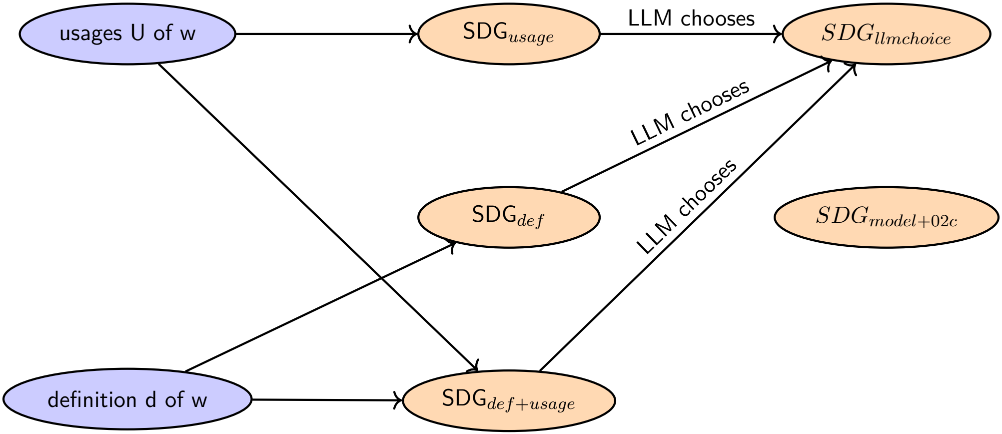
SDG: Models

Approache	Definition Proposal	Retrieved Usages
SDG_{def}	yes	no
SDG_{usage}	no	yes
$SDG_{def+usage}$	yes	yes

SDG: Models

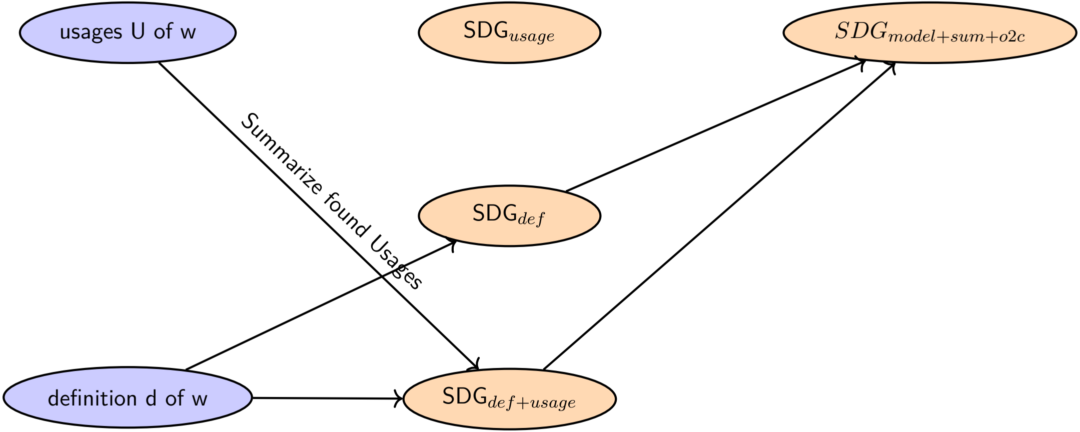

usages U of w
SDG_{usage}
$SDG_{llmchoice}$
SDG_{def}
ELSE
$SDG_{model+02c}$
IF usages available
definition d of w
$\text{SDG}_{def+usage}$

SDG: Models



usages U of w
SDG_{usage}
LLM chooses
$SDG_{llmchoice}$
LLM chooses
LLM chooses
SDG_{def}
$SDG_{model+02c}$
definition d of w
$\text{SDG}_{def+usage}$

SDG: Models


usages U of w
SDG$_{usage}$
$SDG_{model+sum+o2c}$
Summarize found Usages
SDG$_{def}$
definition d of w
SDG$_{def+usage}$

SDG: How to?

- Use Large Language Model: **Gemma (google/gemma-3-12b-it)**
  - Very Large Context length (128k tokens)
- Focus on prompt engineering methods
  - CoT: Chain of Thought
    - Show steps to follow
    - Read inputs, understand domain, improve definitions
  - Retrieve existing definitions
    - Wordnet definitions for headword w
    - O2C trained on wordnet
    - wordnet definitions as referenc
  - Role-based prompting
    - Make the model act as a expert in the field

SDG: Evaluation

How to evaluate?

- **TSV Evaluation:**
    - **T**arget **S**ense **V**erification
    - TSV=WSD [3]
    - WSD Model decides if the sense definition fits the given usage
    - Calculate Average Precision to compare
- **WSI Evaluation:**
    - Can clustering be enhanced using SDG?
    - Basic WSI Model vs. WSD+SDG
    - Calculate Average Adjusted Rand Index for clusters

SDG: TSV Task Description

Input:

- headword w
- proposal of sense definition d for a sense s
- retrieved usage u from U_w

Output:

- TRUE if s with definition d fits usage u
- FALSE else

SDG: TSV Input

Dictionary:

Sense ID	**Definition**
cell 1	*Biological cell*
cell 2	*Cell phone*
cell 3	*Prison cell*

Usages:

Context	**Sense ID (gold)**
I'm in a cell.	cell 3
My android cell	cell 2
A onion cell	cell 1

SDG: TSV Step

Sense ID	Definition
cell 1	*Biological cell*
cell 2	*Cell phone*
cell 3	*Prison cell*

,

Context	Sense ID (gold)
I'm in a cell.	cell 3
My android cell	cell 2
A onion cell	cell 1

cell 1:

Context	TSV Label
I'm in a cell.	0
My android cell	0
A onion cell	1

SDG: TSV Results

Average Precision of TSV evaluation:

Model	**Pilot**$_{suggestions}$	**FEWS**$_{train-ext}$
Baseline	0.15907	0.12435
$SDG_{llmchoice+o2c}$	0.14757	0.10854
$SDG_{model+o2c}$	0.18224	0.11949
$SDG_{model+goldusages}$	0.16891	**0.12477**
$SDG_{model+sum+o2c}$	**0.18559**	0.09114

TSV Distribution:

TSV Label	**Pilot**$_{suggestions}$	**FEWS**$_{train-ext}$
True (1)	78	7985
False (0)	616	146402

SDG: TSV on Dev3

Average Precision of TSV evaluation on Dev3

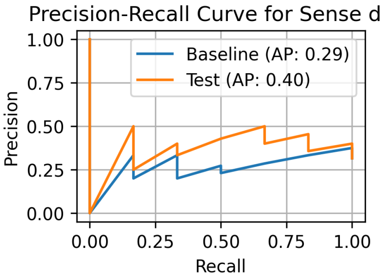
Model	Suggested	Existing
Baseline	0,078	0,252
$SDG_{model+02c}$	**0,118**	**0,280**
$SDG_{model+sum+02c}$	0,096	0,262
$SDG_{llmchoice}$	0,079	0,262

TSV Distribution:

TSV Label	Suggested	Existing
True (1)	752	1481
False (0)	8329	11009

SDG: TSV Results Example DC

Sense ID	Existing Gloss	$SDG_{model+sum+o2c}$
dc-2	District of Columbia as in Washington DC	Washington, D.C. as in the capital district of the United States.



SDG: TSV Results Example DC

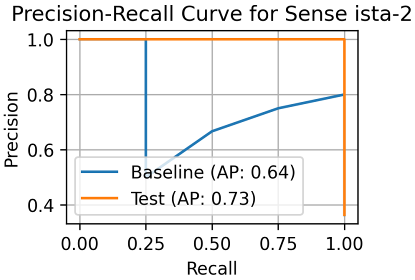
Sense ID	Existing Gloss	$SDG_{model+sum+o2c}$
dc-2	District of Columbia as in Washington DC	Washington, D.C. as in the capital district of the United States.

Helpful usage:

No usage contains: *the capital district of United States*
$\rightarrow$ Model training knowledge has been used here

SDG: TSV Results Example ISTA

Sense ID	Existing Gloss	$SDG_{model+sum+o2c}$
ista-2	Institute of Science and Technology Australia (ISTA), an australian research institute	Institute of Science and Technology Australia (ISTA), an Austrian research institute conducting research in neuroscience, physics, and astrophysics.



SDG: TSV Results Example ISTA

Sense ID	Existing Gloss	$SDG_{model+sum+o2c}$
ista-2	Institute of Science and Technology Australia (ISTA), an australian research institute	Institute of Science and Technology Australia (ISTA), an Austrian research institute conducting research in neuroscience, physics, and astrophysics.

Helpful usage:

... , said the Institute of Science and Technology Austria (ISTA) on Thursday ...

SDG: WSI Task Description

Input:

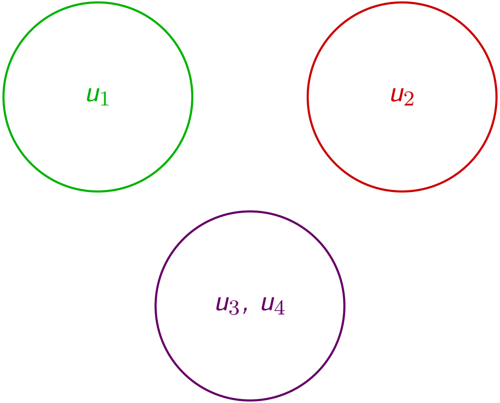
- headword w
- retrieved usages $U_w = \{u_1, u_2, ...\}$ for w

Output:

- a set of sense clusters $C = \{c_1, c_2, ...\}$
- mappings $M : U_w \to C$, assigning each usage $u \in U_w$ to exactly one cluster $c \in C$
- mappings $P : C \to D'$, assigning exactly one cluster $c \in C$ to each generated definition $d' \in D'$

SDG: WSI Clustering example

Usages:



SDG: WSI Input

Dictionary:

Sense ID	Definition
cell 1	*Biological cell*
cell 2	*Cell phone*
cell 3	*Prison cell*

Usages:

Context	Sense ID (gold)
I'm in a cell.	cell 3
My android cell	cell 2
A onion cell	cell 1

SDG: WSI Input

Dictionary:

Sense ID	**Definition**
cell 1	*Biological cell*
cell 2	*Cell phone*
cell 3	*Prison cell*

Usages:

Context	**Sense ID (gold)**
I'm in a cell.	cell 3
My android cell	cell 2
A onion cell	cell 1

SDG: WSI Clustering (correct)

Usages:

I'm in a cell.

My android cell

A onion cell

SDG: WSI Clustering (wrong)

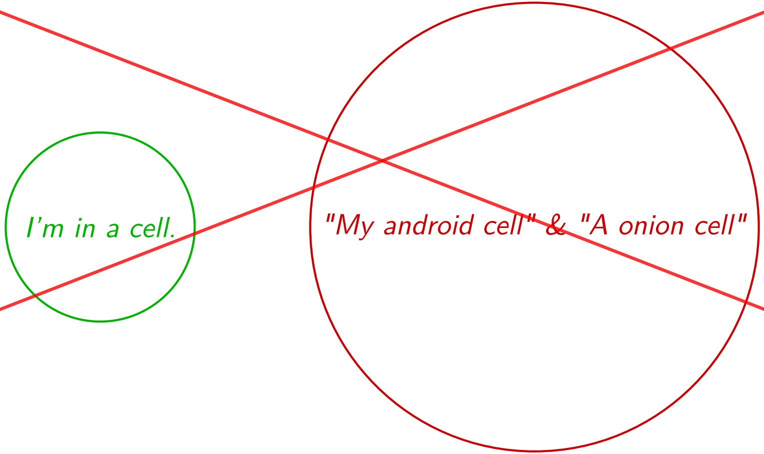
Usages:

I'm in a cell.

"My android cell" & "A onion cell"

SDG: WSI Clustering (wrong)

Usages:



SDG: WSI Steps

1. Run O2C for WSI clusters
2. Run SDG_{usage} on found clusters of each lemma
3. Compare using (average) adjusted rand index

SDG: WSI Results

Average Adjusted Rand Index:

Model	$Pilot$	$FEWS_{train-ext}$	$Dev3$
$Baseline(WSI_{O2C})$	0.16667	0.66389	0.286
$SDG + WSD_{O2C}$	**0.47460**	**0.69247**	**0.318**

Thank you!

References

Denis Kokosinskii, Mikhail Kuklin, and Nikolay Arefyev. *Deep-change at AXOLOTL-24: Orchestrating WSD and WSI Models for Semantic Change Modeling*. https://arxiv.org/abs/2408.05184, 2024.

Mariia Fedorova, Timothee Mickus, Niko Partanen, Janine Siewert, Elena Spaziani, and Andrey Kutuzov. *AXOLOTL'24 Shared Task on Multilingual Explainable Semantic Change Modeling*. In *Proceedings of the 5th Workshop on Computational Approaches to Historical Language Change*, pages 72–91.. Association for Computational Linguistics, Bangkok, Thailand, August 2024. https://aclanthology.org/2024.lchange-1.8.

Bradley Hauer and Grzegorz Kondrak *WiC = TSV = WSD: On the Equivalence of Three Semantic Tasks*. https://arxiv.org/abs/2107.14352