APODICTUS

Automatic Processing Of DICTionary Update candidateS

Felix Blessing[1] Johannes S. Sax[1] Julian Kaufmann[1]
Wei Zhao[3] Nikolay Arefyev[2] Dominik Schlechtweg[1]
first.last@{ims.uni-stuttgart.de/abdn.ac.uk}
University of Stuttgart[1], University of Oslo[2], University of Aberdeen[3]

March 2026

Motivation

- Language is **constantly evolving**

word	PoS	sense description
spam	noun	a tinned meat product made mainly from ham
spam	verb	send the same message indiscriminately to (a large number of internet users)
spam	verb	To press or strike (a computer key, button, etc.) many times in quick succession

Table: Set of dictionary entries of "spam"

Implications for lexicographers:

→ Need to continuously monitor the use of language and update dictionary

Motivation

Oxford University Press maintains various dictionaries, including the *Oxford English Dictionary*.

Their approach:

- Maintain an internal database of update candidates (**LEMUR**), constantly updated with data from diverse sources.
- Decide for each entry: *Should it be added?*
  - Conduct research on the entry.

Data: Dictionary

sense_id	lemma	PoS	gloss
spam_01	spam	noun	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_02	spam	noun	a tinned meat product made mainly from ham
spam_03	spam	verb	send the same message indiscriminately to (a large number of internet users)
...	...	...	...

Table: Example dictionary

Data: Update Candidate Database (LEMUR)

sense_id	lemma	PoS	sense description
LMR2-81764	spam	verb	Slang. To press or strike (a computer key, button, etc.) many times in quick succession.
LMR2-42341	abstrahible	adjective	Abstractable
...	...	...	...

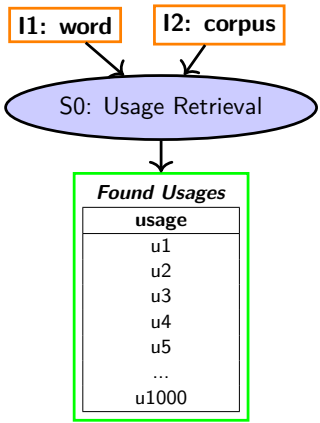
Table: Update candidate examples from LEMUR

Goal of our Work: Automate Process

sense_id	lemma	PoS	gloss	score
LMR2-81764	spam	verb	Slang. To press or strike (a computer key, button, etc.) many times in quick succession.	0.623
LMR2-42341	abstrahible	adjective	Abstractable	0.002
...	...	...	...	...

Table: Update candidate examples from LEMUR

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: 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: Evaluation

On usages from retrieval run that included 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

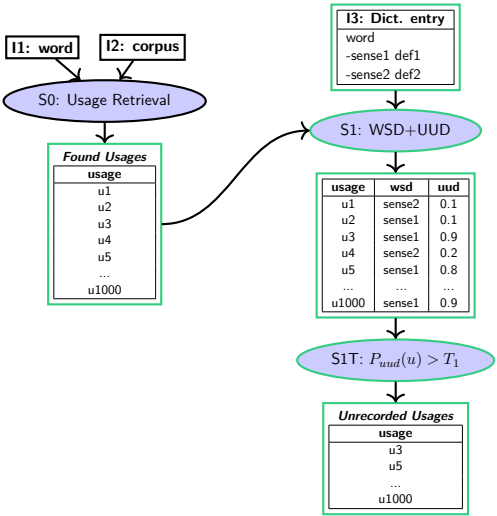
S0: Evaluation

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%

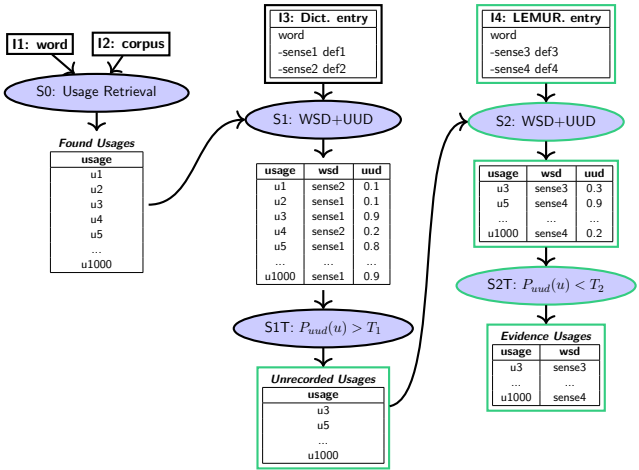
> unforgettable hook and the video is widely shared. Perhaps, with our goldfish memory, we will soon forget about the angry don
>
> NOW-2311IN-103330153-30817188830

S1: Filter Recorded Usages



- Use dictionary to remove usages that contain known/recorded senses
- **W**ord **S**ense **D**isambiguation: Assign to each usage the closest sense
- **U**nknown **U**sage **D**etection: Score $[0, 1]$ for likelihood of usage containing an unrecorded sense, based on distance between WSD sense and usage

S2: Find LEMUR evidence



- Repeat procedure, compare against LEMUR update candidates
- Label usages as:
  - **recorded**: evidence for assigned LEMUR sense
  - **unrecorded**: sense unknown in dictionary and LEMUR

Output

lemma	sense_id	total_usages	evidence_count	frequency	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
- **frequency** = $\frac{\text{evidence\_count}}{\text{total\_usages}}$

List of evidence usages:

sense_id	lemma	usage
LMR2-81764	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
LMR2-81764	spam	...

Table: Usages assigned to the LEMUR sense proposal.

Evaluation: Test Dataset

- Annotate usages of 24 in-dict and 24 out-of-dict words
  - **in-dict**: Target word has a dictionary entry with other recorded senses
  - **out-of-dict**: New word, no dictionary entry
- Annotate usages with the sense they contain
- 2 external annotators, both native english speakers

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

Table: Dataset structure.

Evaluation: USD Models

- **Baseline USD Models (distance-based)**
  - Euclidean distance
  - Manhattan distance
  - ...
- **Logistic Regression USD Models**
  - Russian USD Weights
  - Finnish USD Weights
  - Own USD Weights

Evaluation

- For 10.000 combinations of Thresholds T_1 and T_2, evaluate:
  - Macro-Precision
  - Coverage: Number of words with at least one evidence usage found

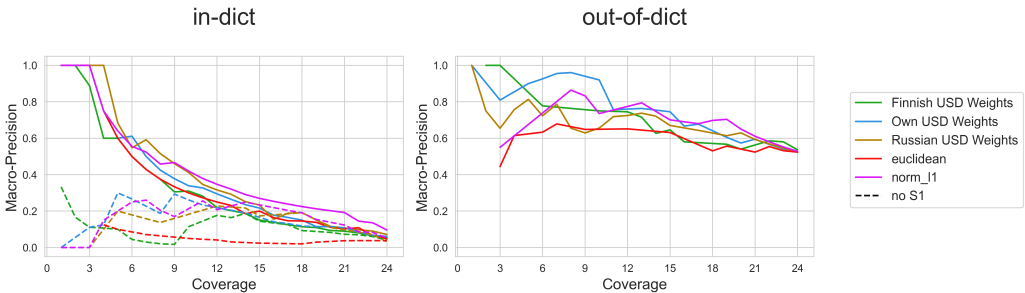
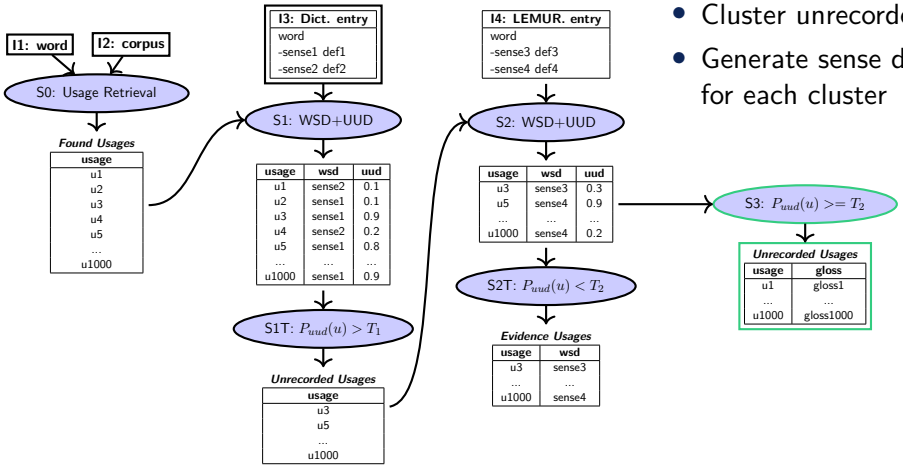


Figure: Maximum precision achievable for given coverage.

S3: Sense Proposal Generation



- Cluster unrecorded usages
- Generate sense definition proposal for each cluster

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