AXOLOTL'24 Shared Task on Multilingual Explainable Semantic Change Modeling

Mariia Fedorova[1], Timothee Mickus[2], Niko Partanen[2], Janine Siewert[2], Elena Spaziani[3], Andrey Kutuzov[1]

[1]University of Oslo, Norway; [2]University of Helsinki, Finland; [3]Sapienza University of Rome, Italy

[1]{mariiaf, andreku}@ifi.uio.no; [2]firstname.lastname@helsinki.fi; [3]elena.spaziani@uniroma1.it

5th International Workshop on Computational Approaches to Historical Language Change (LChange'24)
at ACL 2024, Bangkok, Thailand

Motivation

Why just another LSCD shared task?

DIACR-Ita @ EVALITA2020: Overview of the EVALITA2020 Diachronic Lexical Semantics (DIACR-Ita) Task

Pierpaolo Basile
Dept. of Computer Science
University of Bari, Italy
pierpaolo.basile@uniba.it

Annalina Caputo
ADAPT Centre
School of Computing, Dublin City University
annalina.caputo@dcu.ie

Tommaso Caselli
CLCG
University of Groningen, Netherlands
t.caselli@rug.nl

Pierluigi Cassotti
Dept. of Computer Science
University of Bari, Italy
pierluigi.cassotti@uniba.it

Rossella Varvara
DILEF
University of Florence, Italy
rossella.varvara@unifi.it

RuShiftEval: a shared task on semantic shift detection for Russian

Pivovarova Lidia
University of Helsinki
Finland
lidia.pivovarova@helsinki.fi

Kutuzov Andrey
University of Oslo
Norway
andreku@ifi.uio.no

SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection

Dominik Schlechtweg,[♣] Barbara McGillivray,[◇,♡] Simon Hengchen,[♠,∘] Haim Dubossarsky,[♡] Nina Tahmasebi[♠]
semeval2020lexicalsemantickchange@turing.ac.uk
[♣]University of Stuttgart, [◇]The Alan Turing Institute, [♡]University of Cambridge
[♠]University of Gothenburg

LSCDiscovery: A shared task on semantic change discovery and detection in Spanish

Frank D. Zamora-Reina[1], Felipe Bravo-Marquez[1], Dominik Schlechtweg[2]
[1]Department of Computer Science, University of Chile, IMFD & CENIA
[2]Institute for Natural Language Processing, University of Stuttgart
fzamora@dcc.uchile.cl, fbravo@dcc.uchile.cl,
schlecdk@ims.uni-stuttgart.de

Contents

Shared task overview

February 2024 – April 2024

Ascertain and e**X**plain **O**verhauls of the **L**exicon **O**ver **T**ime at **L**Change'24

The task: **update a dictionary**

- **Subtask 1**: Given target word senses from an old time period and target word usages from a new time period, assign old or newly gained senses to the usages from the new time period;
- **Subtask 2**: Provide definitions of the gained senses

The participants could participate in both tasks, or in one of them only

The task

Data sources

- **Finnish**: Dictionary of Old Literary Finnish Institute for the Languages of Finland (2023); old 1543–1699, new 1700–1810
- **Russian**: Dal's Explanatory Dictionary of the Living Great Russian Language Dal (1909) (old, XIX century) and Wiktionary-based CODWOE Mickus et al. (2022) (new, modern)
- **German**: DWUG DE Sense dataset Schlechtweg (2023); old, XIX century and new 1946-1990

Statistics

Number of samples in AXOLOTL'24 splits

Language	Period	Train	Dev	Test
Finnish	New	47242	3351	3264
	Old	45897	3203	3461
	Total	93139	6554	6725
Russian	New	4581	1605	1702
	Old	1912	421	424
	Total	6493	2026	2126
German	New	—	—	568
	Old	—	—	584
	Total	—	—	1152

Number of target words in AXOLOTL'24 splits

Language	Train	Dev	Test
Finnish	4289	254	275
Russian	924	201	211
German	—	—	24

Data annotation

Finnish

- detection of a target word position in an example by Levenshtein distance
- manual verification of the target word positions in the validation and test splits
- edge cases: punctuation, parts of compound words

Russian

- mapping between Dal senses and CoDWoE senses by decision tree classifier trained on cosine similarity between sentence-transformers Reimers and Gurevych (2020) embeddings of definitions (for 228 manually annotated definition pairs)
- manual verification of the mapping
- edge cases: CoDWOe has more granular senses (one sense in Dal may correspond to many senses in CoDWoE; some Dal samples contain definitions only and no examples)

Contents

Subtask 1

Subtask 1 example

Russian target word экспресс:

Provided sense	Usage	Correct sense
1, old	**express** train, especially fast, express	1, old
?	I was traveling by an **express** train, in a sleeping car	1, old
?	But the other client of this bookmaker was unlucky. He placed 700 thousand rubles on **a combined bet**, in which he included a bet on "Lyon" with betting odds (0)	2, new
?	In this night **train**, which distinguished itself from all other trains by its pre-war comfort...	1, old
?	write in details what you see and send it to me in Otradnoe by **express mail**	3, new

Evaluation

- Adjusted Rand Index (ARI) for all predictions (senses, assigned to usages from the new time period); measures performance in WSI

Subtask 1

Subtask 1 example

Russian target word экспресс:

Provided sense	Usage	Correct sense
1, old	**express** train, especially fast, express	1, old
?	I was traveling by an **express** train, in a sleeping car	1, old
?	But the other client of this bookmaker was unlucky. He placed 700 thousand rubles on **a combined bet**, in which he included a bet on "Lyon" with betting odds (0)	2, new
?	In this night **train**, which distinguished itself from all other trains by its pre-war comfort...	1, old
?	write in details what you see and send it to me in Otradnoe by **express mail**	3, new

Evaluation

- Adjusted Rand Index (ARI) for all predictions (senses, assigned to usages from the new time period); measures performance in WSI
- macro-F1 for all predictions, where usages from the new time period had senses previously existing in the old time period as the gold answers; measures performance in WSD

Contents

Subtask 2

Subtask 2 example

The target word экспресс

Sense	Usages	Gold definition
1, old	I was traveling by an **express** train, in a sleeping car	means of transport (train, ship, bus etc.), traveling at an increased speed, stopping only at major stations
1, old	In this night **express**, which distinguished itself from all other trains by its pre-war comfort...	means of transport (train, ship, bus etc.), traveling at an increased speed, stopping only at major stations
2, new	But the other client of this bookmaker was unlucky. He placed 700 thousand rubles on **a combined bet**, in which he included a bet on "Lyon" with betting odds (0)	special. bet on several independent outcomes
3, new	write in details what you see and send it to me in Otradnoe by **express** mail	colloquial. express mail

Evaluation

BLEU/ROUGE and BERTScore. The final score is averaged across target words

Contents

Our baselines

Subtask 1 baseline

Affinity Propagation Frey and Dueck (2007) on sentence embeddings using LEALLA-large model Mao and Nakagawa (2023)

Subtask 2 baseline

- fine-tuning multilingual causal language model XGLM Lin et al. (2021) as a Siamese network
- sentence-level representations of usage examples are obtained by pooling XGLM's output embeddings and applying a learned linear projection
- using the sentence embeddings obtained in the previous step as an input to the same XGLM to generate the lexicographic definition

Subtask 1 (6 teams)

Subtask 1 results (best submissions per team by averaged Fi-Ru-De)

Team	Fi-Ru-De	Fi-Ru	Fi	Ru	De
Deep-change	**41.3**	**34.9**	**63.8**	05.9	**54.3**
Holotniekat	31.2	32.0	59.6	04.3	29.8
TartuNLP	31.0	26.8	43.7	09.8	39.6
IMS_Stuttgart	28.7	27.4	54.8	00.0	31.4
ABDN-NLP	22.1	28.1	55.3	00.9	10.2
WooperNLP	18.7	28.0	42.8	**13.2**	00.0
Baseline	04.1	05.1	02.3	07.9	02.2
Random sense baseline	19.0	26.3	52.2	00.4	04.4
MFS baseline	24.2	30.1	59.6	00.5	12.5

ARI $\times 100$

Team	Fi-Ru-De	Fi-Ru	Fi	Ru	De
Deep-change	**75.0**	**75.3**	**75.6**	**75.0**	**74.5**
Holotniekat	64.1	65.8	65.5	66.1	60.8
TartuNLP	59.0	59.5	55.0	64.0	58.0
ABDN-NLP	48.7	58.0	59.0	57.0	30.0
IMS_Stuttgart	43.1	32.8	65.5	00.0	63.8
WooperNLP	31.6	47.5	50.3	44.6	00.0
Baseline	20.7	24.5	23.0	26.0	13.0
Random sense baseline	53.3	59.9	62.1	57.7	40.1
MFS baseline	52.6	61.6	65.4	57.7	34.7

F1 $\times 100$

Subtask 1 (6 teams)

Subtask 1 results (best submissions per team by averaged Fi-Ru-De)

Team	Fi-Ru-De	Fi-Ru	Fi	Ru	De
Deep-change	**41.3**	**34.9**	**63.8**	05.9	**54.3**
Holotniekat	31.2	32.0	59.6	04.3	29.8
TartuNLP	31.0	26.8	43.7	09.8	39.6
IMS_Stuttgart	28.7	27.4	54.8	00.0	31.4
ABDN-NLP	22.1	28.1	55.3	00.9	10.2
WooperNLP	18.7	28.0	42.8	**13.2**	00.0
Baseline	04.1	05.1	02.3	07.9	02.2
Random sense baseline	19.0	26.3	52.2	00.4	04.4
MFS baseline	24.2	30.1	59.6	00.5	12.5

ARI $\times 100$

Team	Fi-Ru-De	Fi-Ru	Fi	Ru	De
Deep-change	**75.0**	**75.3**	**75.6**	**75.0**	**74.5**
Holotniekat	64.1	65.8	65.5	66.1	60.8
TartuNLP	59.0	59.5	55.0	64.0	58.0
ABDN-NLP	48.7	58.0	59.0	57.0	30.0
IMS_Stuttgart	43.1	32.8	65.5	00.0	63.8
WooperNLP	31.6	47.5	50.3	44.6	00.0
Baseline	20.7	24.5	23.0	26.0	13.0
Random sense baseline	53.3	59.9	62.1	57.7	40.1
MFS baseline	52.6	61.6	65.4	57.7	34.7

F1 $\times 100$

- Deep-change: GlossReader, no novel senses; Kokosinskii et al. (2024)

Subtask 1 (6 teams)

Subtask 1 results (best submissions per team by averaged Fi-Ru-De)

Team	Fi-Ru-De	Fi-Ru	Fi	Ru	De
Deep-change	**41.3**	**34.9**	**63.8**	05.9	**54.3**
Holotniekat	31.2	32.0	59.6	04.3	29.8
TartuNLP	31.0	26.8	43.7	09.8	39.6
IMS_Stuttgart	28.7	27.4	54.8	00.0	31.4
ABDN-NLP	22.1	28.1	55.3	00.9	10.2
WooperNLP	18.7	28.0	42.8	**13.2**	00.0
Baseline	04.1	05.1	02.3	07.9	02.2
Random sense baseline	19.0	26.3	52.2	00.4	04.4
MFS baseline	24.2	30.1	59.6	00.5	12.5

ARI $\times 100$

Team	Fi-Ru-De	Fi-Ru	Fi	Ru	De
Deep-change	**75.0**	**75.3**	**75.6**	**75.0**	**74.5**
Holotniekat	64.1	65.8	65.5	66.1	60.8
TartuNLP	59.0	59.5	55.0	64.0	58.0
ABDN-NLP	48.7	58.0	59.0	57.0	30.0
IMS_Stuttgart	43.1	32.8	65.5	00.0	63.8
WooperNLP	31.6	47.5	50.3	44.6	00.0
Baseline	20.7	24.5	23.0	26.0	13.0
Random sense baseline	53.3	59.9	62.1	57.7	40.1
MFS baseline	52.6	61.6	65.4	57.7	34.7

F1 $\times 100$

- Deep-change: GlossReader, no novel senses; Kokosinskii et al. (2024)
- WooperNLP: clustering by cosine similarity between sentence embeddings

Subtask 1 (6 teams)

Subtask 1 results (best submissions per team by averaged Fi-Ru-De)

Team	Fi-Ru-De	Fi-Ru	Fi	Ru	De
Deep-change	**41.3**	**34.9**	**63.8**	05.9	**54.3**
Holotniekat	31.2	32.0	59.6	04.3	29.8
TartuNLP	31.0	26.8	43.7	09.8	39.6
IMS_Stuttgart	28.7	27.4	54.8	00.0	31.4
ABDN-NLP	22.1	28.1	55.3	00.9	10.2
WooperNLP	18.7	28.0	42.8	**13.2**	00.0
Baseline	04.1	05.1	02.3	07.9	02.2
Random sense baseline	19.0	26.3	52.2	00.4	04.4
MFS baseline	24.2	30.1	59.6	00.5	12.5

ARI $\times 100$

Team	Fi-Ru-De	Fi-Ru	Fi	Ru	De
Deep-change	**75.0**	**75.3**	**75.6**	**75.0**	**74.5**
Holotniekat	64.1	65.8	65.5	66.1	60.8
TartuNLP	59.0	59.5	55.0	64.0	58.0
ABDN-NLP	48.7	58.0	59.0	57.0	30.0
IMS_Stuttgart	43.1	32.8	65.5	00.0	63.8
WooperNLP	31.6	47.5	50.3	44.6	00.0
Baseline	20.7	24.5	23.0	26.0	13.0
Random sense baseline	53.3	59.9	62.1	57.7	40.1
MFS baseline	52.6	61.6	65.4	57.7	34.7

F1 $\times 100$

- Deep-change: GlossReader, no novel senses; Kokosinskii et al. (2024)
- WooperNLP: clustering by cosine similarity between sentence embeddings
- Holotniekat: clustering by cosine similarity between sentence embeddings Brückner et al. (2024)

Subtask 1 (6 teams)

Subtask 1 results (best submissions per team by averaged Fi-Ru-De)

Team	Fi-Ru-De	Fi-Ru	Fi	Ru	De
Deep-change	**41.3**	**34.9**	**63.8**	05.9	**54.3**
Holotniekat	31.2	32.0	59.6	04.3	29.8
TartuNLP	31.0	26.8	43.7	09.8	39.6
IMS_Stuttgart	28.7	27.4	54.8	00.0	31.4
ABDN-NLP	22.1	28.1	55.3	00.9	10.2
WooperNLP	18.7	28.0	42.8	**13.2**	00.0
Baseline	04.1	05.1	02.3	07.9	02.2
Random sense baseline	19.0	26.3	52.2	00.4	04.4
MFS baseline	24.2	30.1	59.6	00.5	12.5

ARI $\times 100$

Team	Fi-Ru-De	Fi-Ru	Fi	Ru	De
Deep-change	**75.0**	**75.3**	**75.6**	**75.0**	**74.5**
Holotniekat	64.1	65.8	65.5	66.1	60.8
TartuNLP	59.0	59.5	55.0	64.0	58.0
ABDN-NLP	48.7	58.0	59.0	57.0	30.0
IMS_Stuttgart	43.1	32.8	65.5	00.0	63.8
WooperNLP	31.6	47.5	50.3	44.6	00.0
Baseline	20.7	24.5	23.0	26.0	13.0
Random sense baseline	53.3	59.9	62.1	57.7	40.1
MFS baseline	52.6	61.6	65.4	57.7	34.7

F1 $\times 100$

- Deep-change: GlossReader, no novel senses; Kokosinskii et al. (2024)
- WooperNLP: clustering by cosine similarity between sentence embeddings
- Holotniekat: clustering by cosine similarity between sentence embeddings Brückner et al. (2024)
- TartuNLP: GlossBERT with XLM-RoBERTa Dorkin and Sirts (2024)

Subtask 1 (6 teams)

Subtask 1 results (best submissions per team by averaged Fi-Ru-De)

Team	Fi-Ru-De	Fi-Ru	Fi	Ru	De
Deep-change	**41.3**	**34.9**	**63.8**	05.9	**54.3**
Holotniekat	31.2	32.0	59.6	04.3	29.8
TartuNLP	31.0	26.8	43.7	09.8	39.6
IMS_Stuttgart	28.7	27.4	54.8	00.0	31.4
ABDN-NLP	22.1	28.1	55.3	00.9	10.2
WooperNLP	18.7	28.0	42.8	**13.2**	00.0
Baseline	04.1	05.1	02.3	07.9	02.2
Random sense baseline	19.0	26.3	52.2	00.4	04.4
MFS baseline	24.2	30.1	59.6	00.5	12.5

ARI $\times 100$

Team	Fi-Ru-De	Fi-Ru	Fi	Ru	De
Deep-change	**75.0**	**75.3**	**75.6**	**75.0**	**74.5**
Holotniekat	64.1	65.8	65.5	66.1	60.8
TartuNLP	59.0	59.5	55.0	64.0	58.0
ABDN-NLP	48.7	58.0	59.0	57.0	30.0
IMS_Stuttgart	43.1	32.8	65.5	00.0	63.8
WooperNLP	31.6	47.5	50.3	44.6	00.0
Baseline	20.7	24.5	23.0	26.0	13.0
Random sense baseline	53.3	59.9	62.1	57.7	40.1
MFS baseline	52.6	61.6	65.4	57.7	34.7

F1 $\times 100$

- Deep-change: GlossReader, no novel senses; Kokosinskii et al. (2024)
- WooperNLP: clustering by cosine similarity between sentence embeddings
- Holotniekat: clustering by cosine similarity between sentence embeddings Brückner et al. (2024)
- TartuNLP: GlossBERT with XLM-RoBERTa Dorkin and Sirts (2024)
- ABDN-NLP: NeighborClustering Ma et al. (2024)

Subtask 1 (6 teams)

Subtask 1 results (best submissions per team by averaged Fi-Ru-De)

Team	Fi-Ru-De	Fi-Ru	Fi	Ru	De
Deep-change	**41.3**	**34.9**	**63.8**	05.9	**54.3**
Holotniekat	31.2	32.0	59.6	04.3	29.8
TartuNLP	31.0	26.8	43.7	09.8	39.6
IMS_Stuttgart	28.7	27.4	54.8	00.0	31.4
ABDN-NLP	22.1	28.1	55.3	00.9	10.2
WooperNLP	18.7	28.0	42.8	**13.2**	00.0
Baseline	04.1	05.1	02.3	07.9	02.2
Random sense baseline	19.0	26.3	52.2	00.4	04.4
MFS baseline	24.2	30.1	59.6	00.5	12.5

ARI $\times 100$

Team	Fi-Ru-De	Fi-Ru	Fi	Ru	De
Deep-change	**75.0**	**75.3**	**75.6**	**75.0**	**74.5**
Holotniekat	64.1	65.8	65.5	66.1	60.8
TartuNLP	59.0	59.5	55.0	64.0	58.0
ABDN-NLP	48.7	58.0	59.0	57.0	30.0
IMS_Stuttgart	43.1	32.8	65.5	00.0	63.8
WooperNLP	31.6	47.5	50.3	44.6	00.0
Baseline	20.7	24.5	23.0	26.0	13.0
Random sense baseline	53.3	59.9	62.1	57.7	40.1
MFS baseline	52.6	61.6	65.4	57.7	34.7

F1 $\times 100$

- Deep-change: GlossReader, no novel senses; Kokosinskii et al. (2024)
- WooperNLP: clustering by cosine similarity between sentence embeddings
- Holotniekat: clustering by cosine similarity between sentence embeddings Brückner et al. (2024)
- TartuNLP: GlossBERT with XLM-RoBERTa Dorkin and Sirts (2024)
- ABDN-NLP: NeighborClustering Ma et al. (2024)
- IMS_Stuttgart: 1) USD with XL-LEXEME Cassotti et al. (2023) 2) hierarchical flat clustering with cosine similarity

Subtask 2 (3 teams)

Subtask 2 results (best submissions per team by averaged Fi-Ru-De)

Average of BLEU and BERTScore, $\times 100$:

Team	Fi-Ru-De	Fi-Ru	Fi	Ru	De
TartuNLP	**46.7**	**54.1**	35.4	**72.8**	32.0
WooperNLP	34.0	34.6	34.9	34.2	**33.0**
ABDN-NLP	25.3	37.9	**40.7**	35.2	00.0
Baseline	21.8	20.5	21.8	19.1	24.5

Systems' coverage of target words with newly gained senses (percents):

Team	Finnish	Russian	German
TartuNLP	87	86	50
WooperNLP	100	91	100
ABDN-NLP	01	03	—
Baseline	100	100	100

Subtask 2 (3 teams)

Subtask 2 results (best submissions per team by averaged Fi-Ru-De)

Average of BLEU and BERTScore, ×100:

Team	Fi-Ru-De	Fi-Ru	Fi	Ru	De
TartuNLP	**46.7**	**54.1**	35.4	**72.8**	32.0
WooperNLP	34.0	34.6	34.9	34.2	**33.0**
ABDN-NLP	25.3	37.9	**40.7**	35.2	00.0
Baseline	21.8	20.5	21.8	19.1	24.5

Systems' coverage of target words with newly gained senses (percents):

Team	Finnish	Russian	German
TartuNLP	87	86	50
WooperNLP	100	91	100
ABDN-NLP	01	03	—
Baseline	100	100	100

- **TartuNLP**: **GlossBERT** fine-tuned to match definitions from Wiktionary (which was the source of the new time period for Russian) Dorkin and Sirts (2024)

Subtask 2 (3 teams)

Subtask 2 results (best submissions per team by averaged Fi-Ru-De)

Average of BLEU and BERTScore, ×100:

Team	Fi-Ru-De	Fi-Ru	Fi	Ru	De
TartuNLP	**46.7**	**54.1**	35.4	**72.8**	32.0
WooperNLP	34.0	34.6	34.9	34.2	**33.0**
ABDN-NLP	25.3	37.9	**40.7**	35.2	00.0
Baseline	21.8	20.5	21.8	19.1	24.5

Systems' coverage of target words with newly gained senses (percents):

Team	Finnish	Russian	German
TartuNLP	87	86	50
WooperNLP	100	91	100
ABDN-NLP	01	03	—
Baseline	100	100	100

- **TartuNLP**: **GlossBERT** fine-tuned to match definitions from Wiktionary (which was the source of the new time period for Russian) Dorkin and Sirts (2024)
- **WooperNLP**: prompting **GPT3.5**
  https://github.com/t-montes/Axolotl24

Subtask 2 (3 teams)

Subtask 2 results (best submissions per team by averaged Fi-Ru-De)

Average of BLEU and BERTScore, ×100:

Team	Fi-Ru-De	Fi-Ru	Fi	Ru	De
TartuNLP	**46.7**	**54.1**	35.4	**72.8**	32.0
WooperNLP	34.0	34.6	34.9	34.2	**33.0**
ABDN-NLP	25.3	37.9	**40.7**	35.2	00.0
Baseline	21.8	20.5	21.8	19.1	24.5

Systems' coverage of target words with newly gained senses (percents):

Team	Finnish	Russian	German
TartuNLP	87	86	50
WooperNLP	100	91	100
ABDN-NLP	01	03	—
Baseline	100	100	100

- **TartuNLP**: **GlossBERT** fine-tuned to match definitions from Wiktionary (which was the source of the new time period for Russian) Dorkin and Sirts (2024)
- **WooperNLP**: prompting **GPT3.5** `https://github.com/t-montes/Axolotl24`
- **ABDN-NLP**: prompting **GPT3.5** Ma et al. (2024)

Contents

Conclusions

- both subtasks proved to be challenging

Conclusions

- both subtasks proved to be challenging
- systems relying on masked language models specifically fine-tuned on a set of curated sense definitions are most robust across languages and tasks

Conclusions

- both subtasks proved to be challenging
- systems relying on masked language models specifically fine-tuned on a set of curated sense definitions are most robust across languages and tasks
- systems which attempt to infer sense knowledge directly from a large generative LM do not fall far behind (e.g. WooperNLP used it to augment subtask 1 data)

Conclusions

- both subtasks proved to be challenging
- systems relying on masked language models specifically fine-tuned on a set of curated sense definitions are most robust across languages and tasks
- systems which attempt to infer sense knowledge directly from a large generative LM do not fall far behind (e.g. WooperNLP used it to augment subtask 1 data)
- most systems demonstrated good cross-lingual capabilities, being able to produce satisfactory predictions for a surprise language (German) without any training data

Conclusions

- both subtasks proved to be challenging
- systems relying on masked language models specifically fine-tuned on a set of curated sense definitions are most robust across languages and tasks
- systems which attempt to infer sense knowledge directly from a large generative LM do not fall far behind (e.g. WooperNLP used it to augment subtask 1 data)
- most systems demonstrated good cross-lingual capabilities, being able to produce satisfactory predictions for a surprise language (German) without any training data
- the datasets are publicly available

Post-evaluation stage and data

Subtask 1:	**Subtask 2:**	**Code and data repository**
https://codalab.lisn.upsaclay.fr/competitions/18570	https://codalab.lisn.upsaclay.fr/competitions/18572	https://github.com/ltgoslo/axolotl24_shared_task

References I

Brückner, C., Zhang, L., and Pecina, P. (2024). Similarity-based cluster merging for semantic change modeling. In Tahmasebi, N., Montariol, S., Kutuzov, A., Hengchen, S., Alfter, D., Periti, F., and Cassotti, P., editors, *Proceedings of the 5th Workshop on Computational Approaches to Historical Language Change*, Bangkok. Association for Computational Linguistics.

Cassotti, P., Siciliani, L., DeGemmis, M., Semeraro, G., and Basile, P. (2023). XL-LEXEME: WiC pretrained model for cross-lingual LEXical sEMantic changE. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1577–1585, Toronto, Canada. Association for Computational Linguistics.

Dal, V. (1909). Explanatory dictionary of the living great Russian language ed. by Boduen de Kurtene [Tolkovy slovar zhivogo velikorusskogo yazyka, pod red. I. A. Boduena de Kurtene].

References II

Dorkin, A. and Sirts, K. (2024). TartuNLP @ AXOLOTL-24: Leveraging classifier output for new sense detection in lexical semantics. In Tahmasebi, N., Montariol, S., Kutuzov, A., Hengchen, S., Alfter, D., Periti, F., and Cassotti, P., editors, *Proceedings of the 5th Workshop on Computational Approaches to Historical Language Change*, Bangkok. Association for Computational Linguistics.

Frey, B. J. and Dueck, D. (2007). Clustering by passing messages between data points. *Science*, 315(5814):972–976.

Institute for the Languages of Finland (2023). Vanhan kirjasuomen sanakirja [Dictionary of Old Literary Finnish]. Digital resource. Last update 24.11.2023. Accessed 24.11.2023.

References III

Kokosinskii, D., Kuklin, M., and Arefyev, N. (2024). Deep-change at AXOLOTL-24: Orchestrating WSD and WSI models for semantic change modeling. In Tahmasebi, N., Montariol, S., Kutuzov, A., Hengchen, S., Alfter, D., Periti, F., and Cassotti, P., editors, *Proceedings of the 5th Workshop on Computational Approaches to Historical Language Change*, Bangkok. Association for Computational Linguistics.

Lin, X. V., Mihaylov, T., Artetxe, M., Wang, T., Chen, S., Simig, D., Ott, M., Goyal, N., Bhosale, S., Du, J., Pasunuru, R., Shleifer, S., Koura, P. S., Chaudhary, V., O'Horo, B., Wang, J., Zettlemoyer, L., Kozareva, Z., Diab, M. T., Stoyanov, V., and Li, X. (2021). Few-shot learning with multilingual language models. *CoRR*, abs/2112.10668.

Ma, X., Schlechtweg, D., and Zhao, W. (2024). Presence or absence: Are unknown word usages in dictionaries? In Tahmasebi, N., Montariol, S., Kutuzov, A., Hengchen, S., Alfter, D., Periti, F., and Cassotti, P., editors, *Proceedings of the 5th Workshop on Computational Approaches to Historical Language Change*, Bangkok. Association for Computational Linguistics.

References IV

Mao, Z. and Nakagawa, T. (2023). LEALLA: Learning lightweight language-agnostic sentence embeddings with knowledge distillation. In Vlachos, A. and Augenstein, I., editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1886–1894, Dubrovnik, Croatia. Association for Computational Linguistics.

Mickus, T., Van Deemter, K., Constant, M., and Paperno, D. (2022). Semeval-2022 task 1: CODWOE – comparing dictionaries and word embeddings. In Emerson, G., Schluter, N., Stanovsky, G., Kumar, R., Palmer, A., Schneider, N., Singh, S., and Ratan, S., editors, *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1–14, Seattle, United States. Association for Computational Linguistics.

References V

Reimers, N. and Gurevych, I. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Schlechtweg, D. (2023). *Human and computational measurement of lexical semantic change*. PhD thesis, University of Stuttgart.