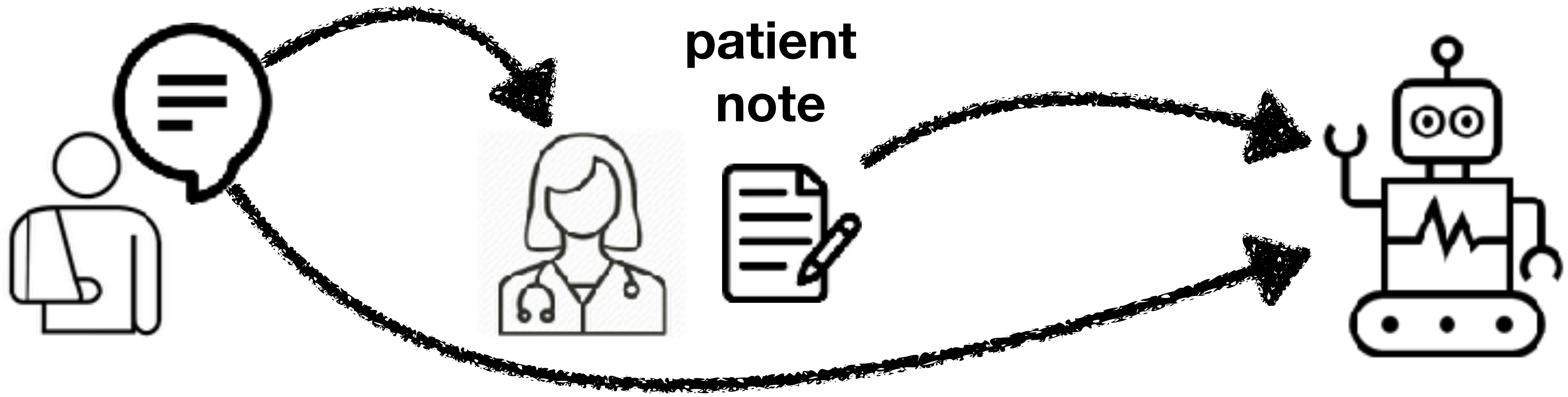


Dutch Language Models & Conversational AI for monitoring functioning and wellbeing

Piek Vossen (piek.vossen@vu.nl)

Computational Linguistics and Text Mining Lab (CLTL)
Vrije Universiteit Amsterdam

Dutch Clinical NLP Workshop, June, 2024, Utrecht

Using NLP for Health

- Building a medical Dutch Large Language Model (LLM)
- Finetuning the medical LLM for text classification: functional levels to patients (A-PROOF)
- Conversational AI for monitoring functioning levels (patient perspective)

Medical note data

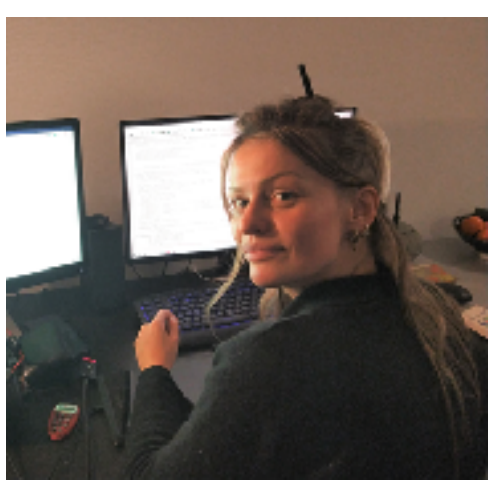
- 2020: Amsterdam UMC keeps over 13 million medical notes from 250K patients (2017, 2018, 2020) on a secure server, expanding the data every year —> currently 20M+

- What knowledge and information is contained in these notes written by professionals besides clinical patient data?

- Reports on observations, conversations, treatments, describing the physical, mental and social conditions of patients over time

- Short (ungrammatical) sentences (10 to 30 per note), expert terminology mixed with common Dutch and English

Dutch Medical Language Model



Stella Verkijk, Master Thesis project, 2021

contact: s.verkijk@vu.nl

model: https://huggingface.co/CLTL/MedRoBERTa.nl

papers: https://scholar.google.com/citations?user=xeMWvjIAAAAJ&hl=nl&oi=ao

Should we use existing (Dutch) Language Models for Medical Text mining?

https://huggingface.co/GroNLP/bert-base-dutch-cased

RobBERT-2023
A Dutch RoBERTa-based Language Model

https://huggingface.co/DTAI-KULeuven/robbert-2023-dutch-large

The language of medical notes

- De stemming ***imponeert*** normofoor, met een normaal modulerend affect
  - Mood impresses normophore, with normal modulating affect.
- Patiënt met ***possibele*** pulmonale aspergillus met oplopend galactomannan onder vori mono
  - Patient with possible pulmonary aspergillus with ascending galactomannan under vori mono).

- Medical jargon
  - Lexical differences
  - Abbreviations
- Different syntax
- Different ambiguity problems
- Typos, spelling mistakes

The language of medical notes

- *get's up and walks to the toilet*
  Patient **loopt** op voor po-stoel en toilet
- *exhausted*
  Patient **loopt** op laatste benen
- *shape*
  (dus "annulus" **loopt** scheef) met daarbij malapositie van de klepbladen
- *got stuck*
  Moeder **loopt** vast in contact
- *falls behind*
  Neu: nomale ontwikkeling, alleen spraak **loopt** wat achter
- *drain runs well*
  infuus **loopt** goed op stand 63
- *quality of walking*
  loopt vlot, normale paslengte en armzwaai, draait vlot.

Why not use medical notes for building a model?

DATA SPECIFICATION

		2017	2018	2020	total
AMC	GB #notes	2.8 2,375,626	3.0 2,451,973	2.0 1,492,573	7.8 6.3M
VuMC	GB #notes	3.0 2,545,515	/	1.5 1,111,116	4.5 3.7M

BUILDING THE MODEL:

PRE-TRAINING STRATEGIES

FROM SCRATCH

Random initialisation, only pre-train on domain specific data and initialise with domain-specific vocabulary directly

EXTENDED PRE-TRAINING

Initialise existing model, in our case RobBERT, and further pre-train on domain-specific data with a domain-specific vocabulary

Medical Language Model

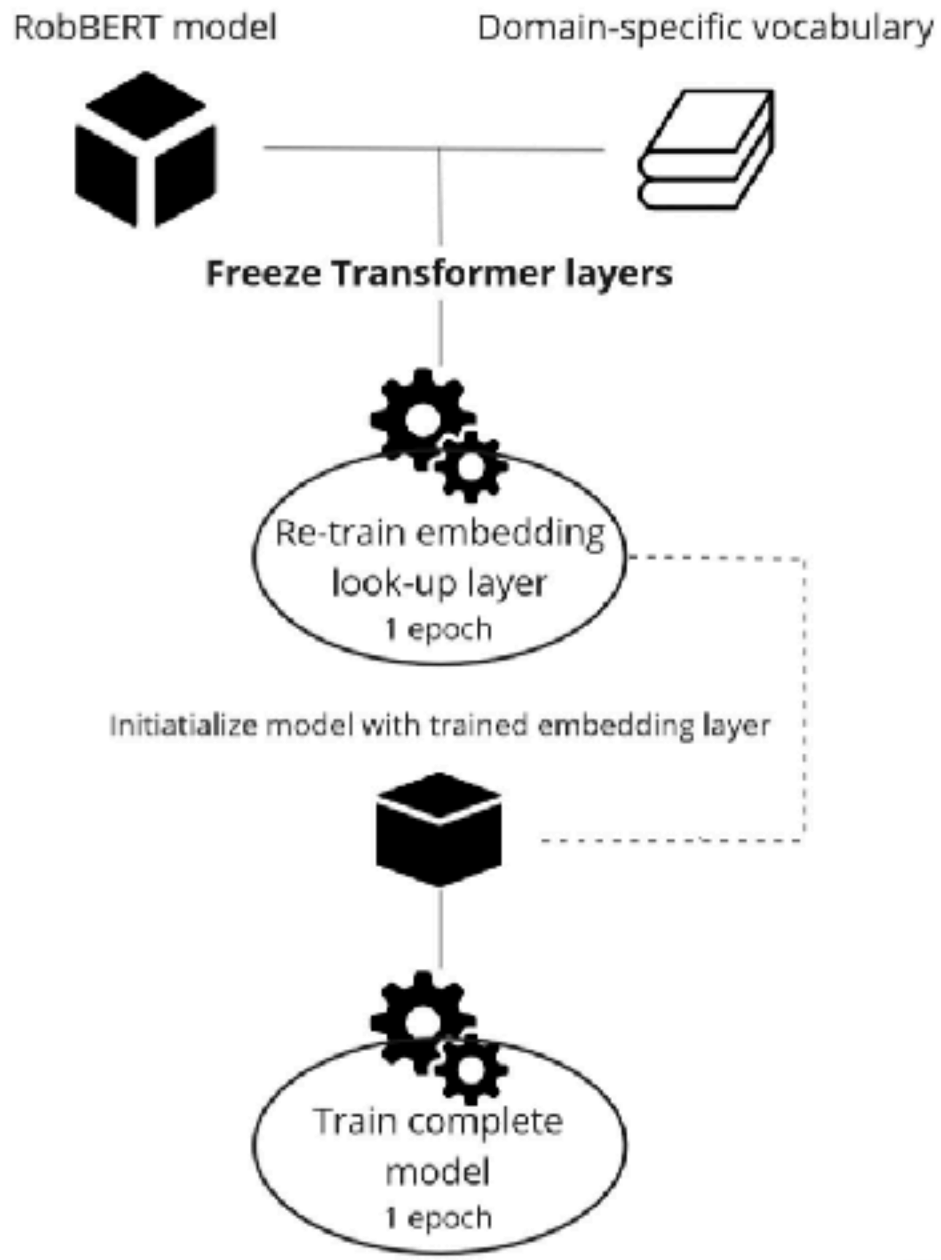


Figure 1: Training process of the extended model

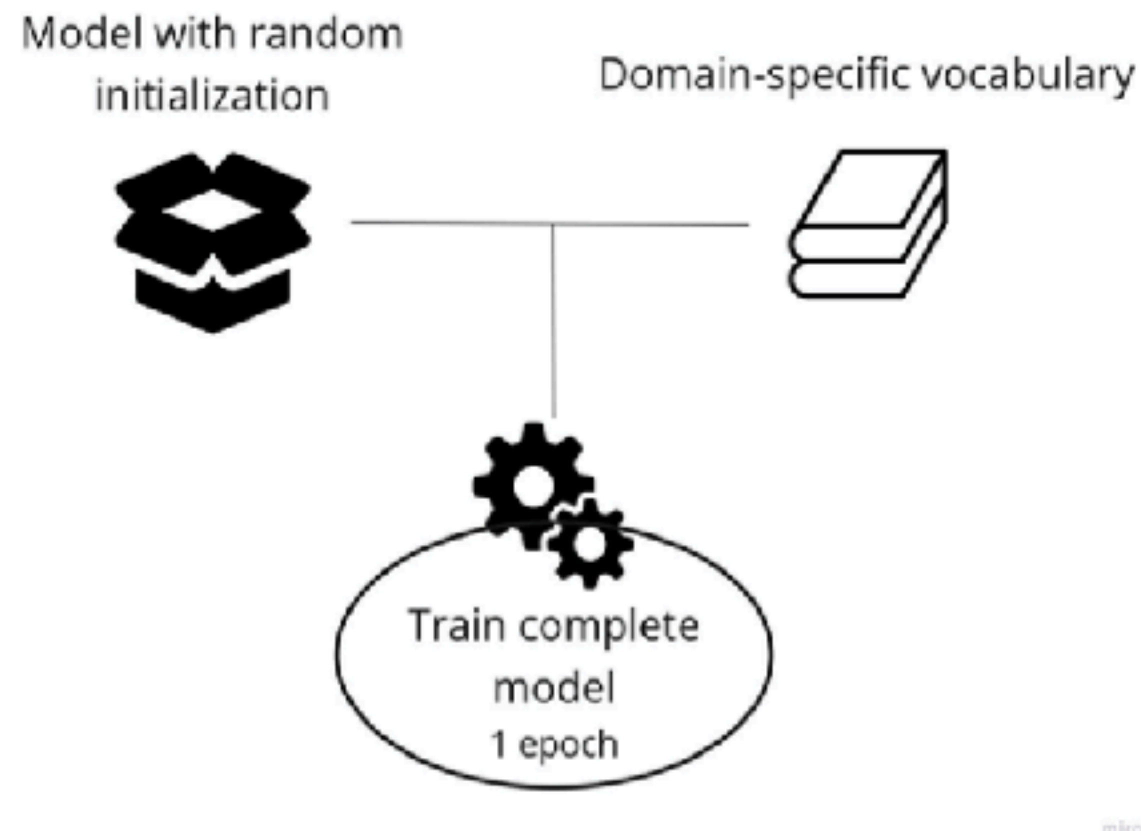


Figure 2: Training from scratch

Evaluation

- Intrinsic evaluation: what is the quality of the text representations of the model —> text similarity (odd-one-out task)

- Extrinsic evaluation: how well can the model be finetuned for specific tasks: Named Entity Recognition, labelling medical notes with WHO's International Classification of Functioning, Disability and Health (ICF)

Intrinsic evaluation: odd-one-out

Sen1	Sen2	Sen3	Annotation
Mobiliseert op de afdeling .	Rolstoelgebonden .	Mob : zelfstandig met de rolstoel .	1
Bemerkt spiermassaverlies (niet geobjectiveerd , verminderde energie , grote stukken lopen gaat niet meer) .	ik merk dat vermoeidheid toeneemt , benauwdheid ook , mn na kleine inspanning en moeheid .	Heeft dan weer het gevoel dat hij kapot is en erg vermoeid .	1
2x per week naar kantoor .	: veeleisende baan	Was daar erg slaperig , erg moe en emotioneel .	3
Zit in de stoel , is erg vermoeid .	Dhr komt nu 2x daags in stoel met passieve lift , daarnaast 1 of 2 keer per dag fysiotherapie ivm vermoeidheid veel verspreiding van intensiteit per dag .	Bemerkt spiermassaverlies (niet geobjectiveerd , verminderde energie , grote stukken lopen gaat niet meer) .	3
Vader denkt dat het lopen beter gaat door de calciumtabletten die PERSON nu slikt .	Is daarna de rest van de dag vermoeid , echter kan nog wel een klein rondje wandelen wat eerst een groter rondje was .	Heeft dan weer het gevoel dat hij kapot is en erg vermoeid .	1

Intrinsic evaluation

- T1 **All** sentences are from the **same** ICF category, but two of them contain overlapping keywords
- T2 **Two** of the sentences are from the **same** ICF category, one is from a different ICF category and the sentences that are from the same category do **not contain overlapping keywords**
- T3 **Two** of the sentences are from the **same** ICF category, one is from a different ICF category and the sentences that are from the same category also **contain at least one overlapping keyword**
- T4 **All** sentences are from the **same** ICF category, but the **functional level** of one sentence **differs** from the functional levels of the other two sentences. Sometimes the difference in functional level is small, other times bigger

ICF categories: level of walking, breathing, eating, mood, energy

	All triples	T1	T2	T3	T4
mBERT	0.57	0.56	0.48	0.68	0.52
BERTje	0.58	0.59	0.49	0.68	0.54
RobBERT	0.57	0.56	0.44	0.68	**0.57**
From Scratch	$\boxed{\mathbf{0.65}}$	**0.65**	**0.56**	$\boxed{\mathbf{0.76}}$	**0.57**
Ext. RobBERT Frozen	0.52	0.55	0.41	0.58	0.53
Ext. RobBERT Final	0.58	0.60	0.48	0.65	0.54
Support	824	194	214	268	148

Table 5: Accuracy per model on complete test set and per triple type

keyword representations more important role

	All triples	**T1**	T2	**T3**	T4
mBERT	0.51	0.44	0.49	0.57	**0.55**
BERTje	0.53	0.47	0.49	0.62	0.50
RobBERT	0.53	0.51	0.41	**0.63**	0.53
From Scratch	$\boxed{\mathbf{0.57}}$	**0.55**	**0.55**	$\boxed{\mathbf{0.63}}$	0.53
Ext. RobBERT Frozen	0.49	0.52	0.41	0.53	0.51
Ext. RobBERT Final	0.51	0.54	0.46	0.54	0.51
Support	824	194	214	268	148

Table 6: Accuracy per model on complete test set and per triple type, with keywords removed

EVALUATING THE MODEL: EXTRINSIC IN-DOMAIN

TASK

Sentence classification:

Detecting & classifying categories from the WHO's International Classification of Functioning, Disability and Health (ICF)

42.4K sentences from 2640 notes in training set; 40.2K sentences from 739 notes in test set

Human (semi-)expert annotators

RESULTS

	RobBERT	BERTje	Ext. RobBERT	MedRoBERTa.nl
Walking	0.62	0.62	0.62	0.65
Emotional functions	0.66	0.69	0.66	0.67
Exercise tolerance	0.42	0.45	0.45	0.45
Work and employment	0.40	0.40	0.39	0.39

Currently over 6k notes and 286K sentences

Few data and low IAA

EVALUATING THE MODEL: EXTRINSIC OUT OF DOMAIN

TASK

Dutch Named Entity Recognition

CoNLL 2002 Dutch news articles (Tjong Kim Sang, 2002)

RESULTS

	P	R	F1
BERTje	0.91	0.92	0.91
RobBERT	0.84	0.85	0.84
extended medical RobBERT	0.64	0.68	0.66
MedRoBERTa.nl	0.68	0.72	0.70

Finetuning MedicalRoberta.nl

for Medical Text Mining

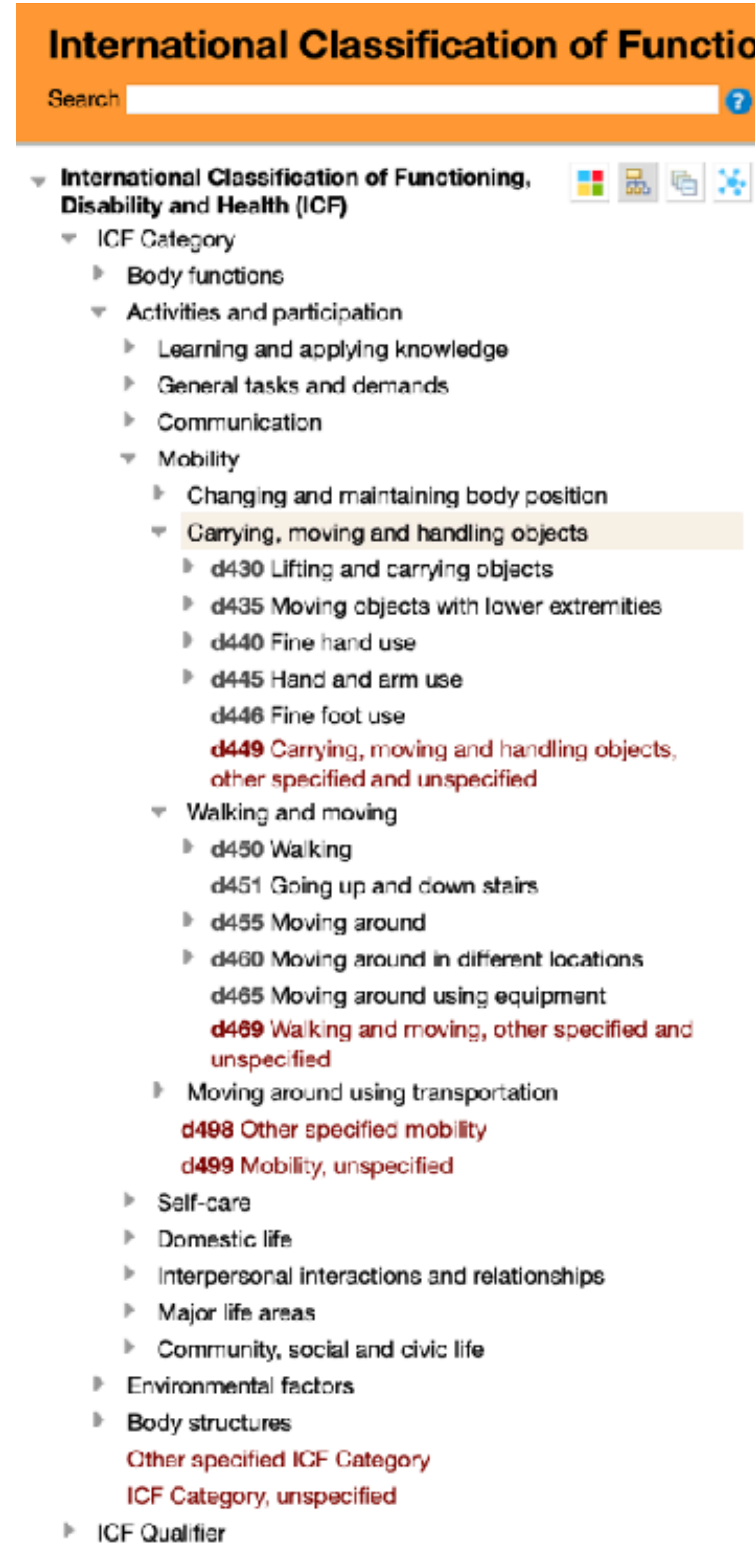
using **implicit** representations

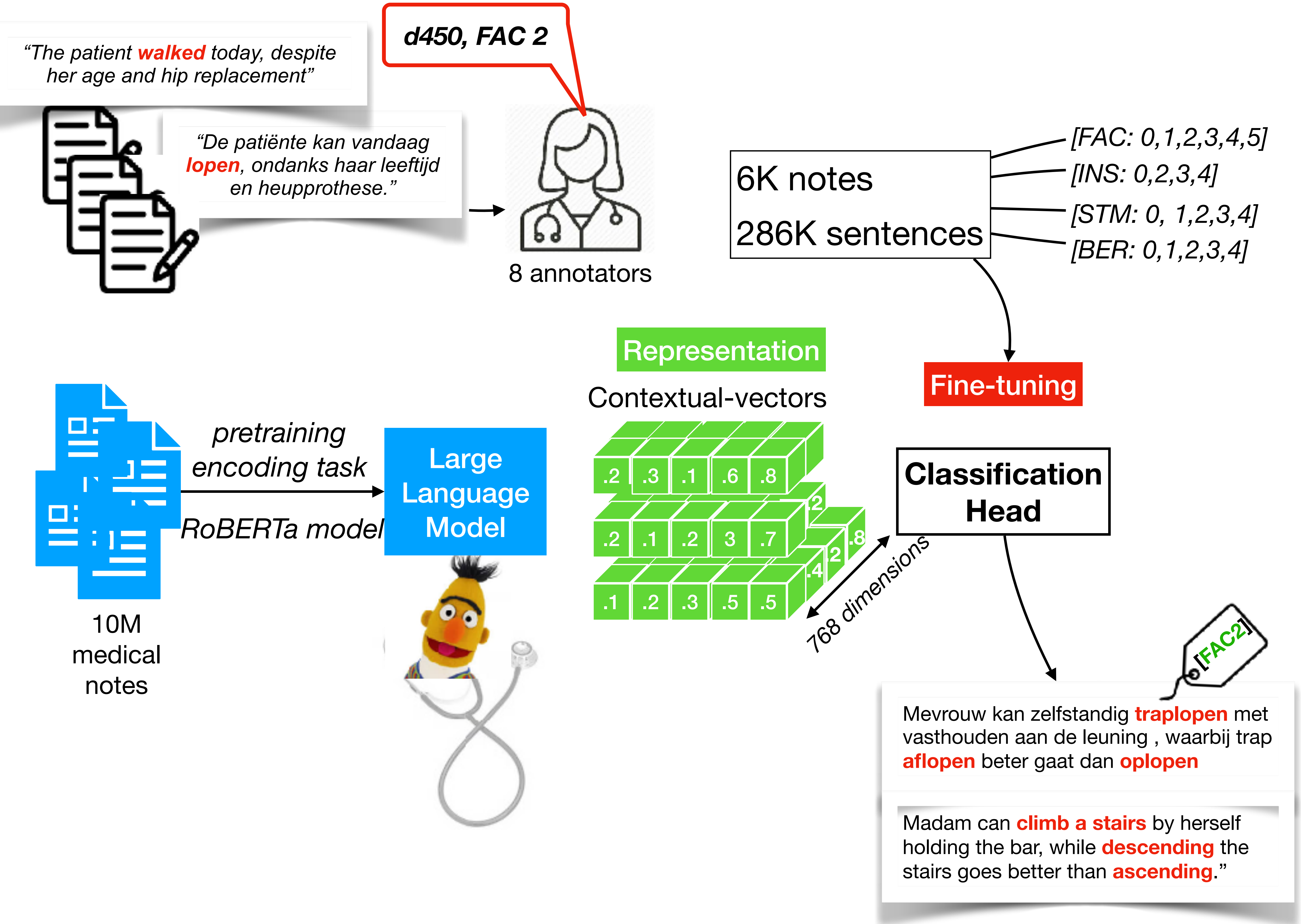
A-PROOF use case

- Clinical treatment of COVID patients:
  - patients experience a broad spectrum of symptoms also after longer clinical treatments
  - it is a new disease about which we still need to learn a lot
  - there is no large-scale structured data on their recovery to make any predictions
- Can we derive such data and knowledge on the recovery from clinical notes?
  - Can we develop AI to read the Dutch notes and generate interpretations
  - Can we use this generated data to learn about the functional recovery over time during and after hospital admission?
  - What are key factors that determine recovery patterns?

Functional level of patients

- WHO International Classification of Functioning, Disability and Health (ICF)
- https://www.who.int/standards/classifications/international-classification-of-functioning-disability-and-health)




"The patient walked today, despite her age and hip replacement"
"De patiënte kan vandaag lopen, ondanks haar leeftijd en heupprothese."
d450, FAC 2
8 annotators
6K notes
286K sentences
[FAC: 0,1,2,3,4,5]
[INS: 0,2,3,4]
[STM: 0, 1,2,3,4]
[BER: 0,1,2,3,4]
Fine-tuning
Representation
Contextual-vectors
pretraining encoding task
RoBERTa model
Large Language Model
10M medical notes
.2 .3 .1 .6 .8
.2 .1 .2 3 .7
.1 .2 .3 .5 .5
768 dimensions
Classification Head
[FAC2]
Mevrouw kan zelfstandig traplopen met vasthouden aan de leuning , waarbij trap aflopen beter gaat dan oplopen
Madam can climb a stairs by herself holding the bar, while descending the stairs goes better than ascending."

Annotated Data

ADM=respiration
ATT=attention
BER=work
ENR=energy level
ETN=eat
FAC=walking
INS=exercise tolerance
MBW=weight
STM=mood

only 5 % of the sentences received a label and is relevant

	ADM	ATT	BER	ENR	ETN	FAC	INS	MBW	STM
train	4,988	247	486	989	2,420	2,489	1,967	755	3,390
dev	411	22	29	105	225	119	127	96	147
test	775	39	54	160	382	253	287	125	181
total	6,174	308	569	1,254	3,027	2,861	2,381	976	3,718

Table 4.2: Domain classification: sentences with labels (positive examples)

	ADM	ATT	BER	ENR	ETN	FAC	INS	MBW	STM
train	2,345	175	381	707	1,416	1,631	1,260	546	1,989
dev	188	17	25	71	128	75	78	71	83
test	231	27	34	92	165	95	116	64	94
total	2,764	219	440	870	1,709	1,801	1,454	681	2,166

Table 4.3: Domain classification: notes with labels (positive examples)

Jenia Kim, Stella Verkijk, Edwin Geleijn, Marike van der Leeden, Carel Meskers, Caroline Meskers, Sabina van der Veen, Piek Vossen, and Guy Widdershoven. Modeling Dutch Medical Texts for Detecting Functional Categories and Levels of COVID-19 Patients. Proceedings of the Language Resources and Evaluation Conference (LREC), 2022, Marseille.

Classification results

	ADM	ATT	BER	ENR	ETN	FAC	INS	MBW	STM
precision	1.0	1.0	0.66	0.96	0.95	0.84	0.95	0.87	0.80
recall	0.89	0.56	0.44	0.70	0.72	0.89	0.46	0.87	0.87
F1-score	0.94	0.71	0.50	0.81	0.82	0.86	0.61	0.87	0.84
support	231	27	34	92	165	95	116	64	94

Table 4.4: Domain classification: evaluation on test set, note-level

	ADM	ATT	BER	ENR	ETN	FAC	INS	MBW	STM
precision	0.98	0.98	0.56	0.96	0.92	0.84	0.89	0.79	0.70
recall	0.49	0.41	0.29	0.57	0.49	0.71	0.26	0.62	0.75
F1-score	0.66	0.58	0.35	0.72	0.63	0.76	0.41	0.70	0.72
support	775	39	54	160	382	253	287	125	181

Table 4.5: Domain classification: evaluation on test set, sentence-level

Jenia Kim, Stella Verkijk, Edwin Geleijn, Marike van der Leeden, Carel Meskers, Caroline Meskers, Sabina van der Veen, Piek Vossen, and Guy Widdershoven. Modeling Dutch Medical Texts for Detecting Functional Categories and Levels of COVID-19 Patients. Proceedings of the Language Resources and Evaluation Conference (LREC), 2022, Marseille.

Regression results

	ADM	ATT	BER	ENR	ETN	FAC*	INS*	MBW	STM
MAE	0.37	1.03	1.49	0.43	0.50	0.66	0.61	0.60	0.68
MSE	0.34	1.47	2.85	0.42	0.47	0.93	0.64	0.56	0.87
RMSE	0.58	1.21	1.69	0.65	0.68	0.96	0.80	0.75	0.93
support	200	21	22	70	123	79	74	41	84

Table 4.14: Levels classification: evaluation results, note-level

	ADM	ATT	BER	ENR	ETN	FAC*	INS*	MBW	STM
MAE	0.48	0.99	1.56	0.48	0.59	0.70	0.69	0.81	0.76
MSE	0.55	1.35	3.06	0.49	0.65	0.91	0.80	0.83	1.03
RMSE	0.74	1.16	1.75	0.70	0.81	0.95	0.89	0.91	1.01
support	421	32	26	100	183	139	136	60	155

Table 4.15: Levels classification: evaluation results, sentence-level

Jenia Kim, Stella Verkijk, Edwin Geleijn, Marike van der Leeden, Carel Meskers, Caroline Meskers, Sabina van der Veen, Piek Vossen, and Guy Widdershoven. Modeling Dutch Medical Texts for Detecting Functional Categories and Levels of COVID-19 Patients. Proceedings of the Language Resources and Evaluation Conference (LREC), 2022, Marseille.

Confusion Matrix

ADM=respiration
ATT=attention
BER=work
ENR=energy level
ETN=eat
FAC=walking
INS=exercise tolerance
MBW=weight
STM=mood

	ADM	ATT	BER	ENR	ETN	FAC	INS	MBW	STM	<none>	support
ADM	368	0	0	0	0	0	0	0	0	9	*377*
ATT	0	14	0	0	0	0	0	0	0	0	*14*
BER	0	0	22	1	0	0	4	0	0	24	*51*
ENR	0	0	0	90	0	0	0	0	0	2	*92*
ETN	0	0	0	0	186	0	0	1	0	15	*202*
FAC	2	0	0	1	1	182	5	0	0	41	*232*
INS	1	0	1	0	0	0	83	0	0	18	*103*
MBW	0	0	0	1	4	0	0	77	0	22	*104*
STM	1	0	1	1	1	0	1	0	138	65	*208*
<none>	403	25	30	66	190	71	194	47	43	0	

Figure 4.2: Domain classification: confusion matrix

Most of the confusion comes from false positives and false negatives with <NONE>

Training Data distribution

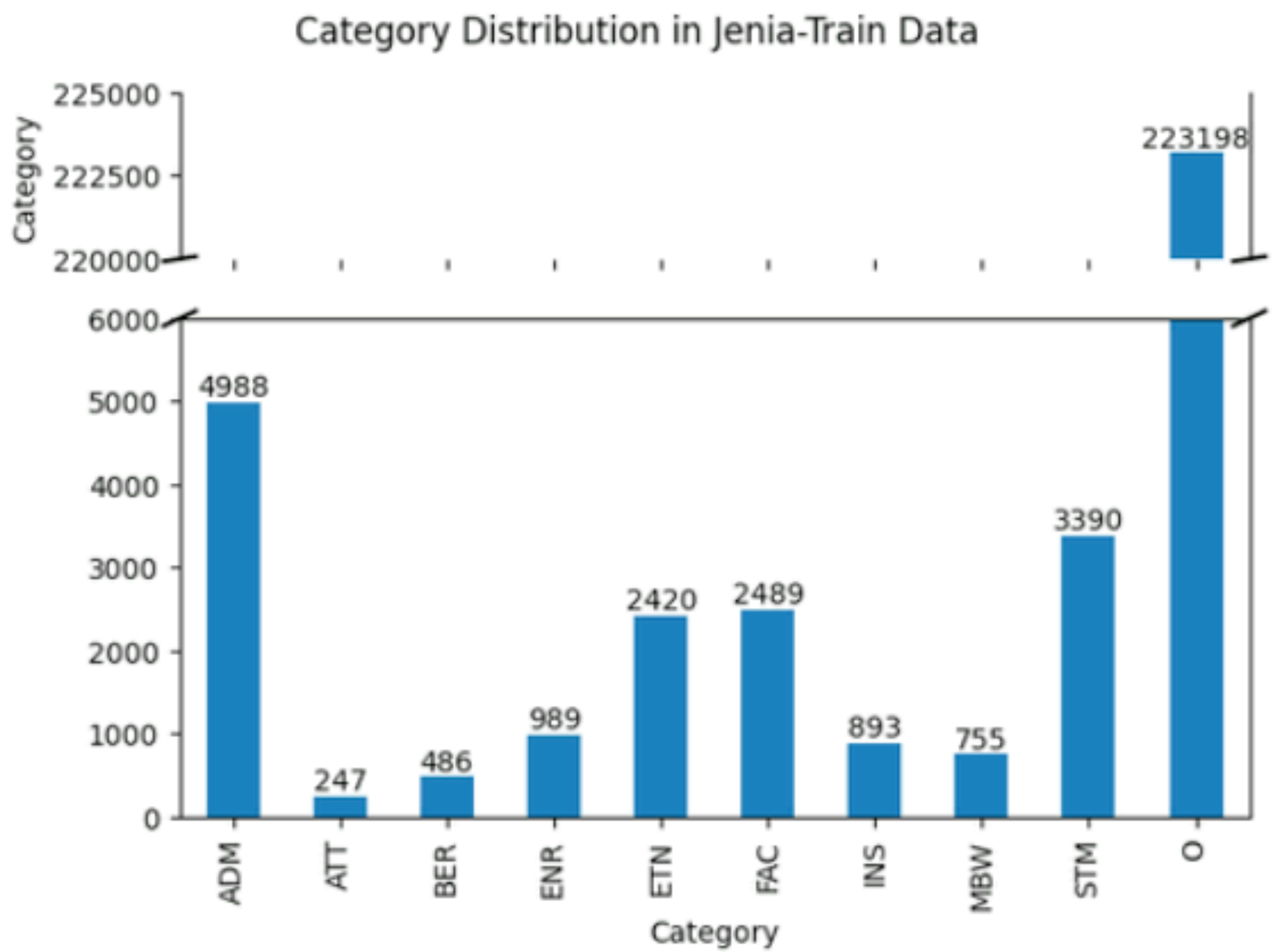


Figure 3.1: Jenia Train Split Category Distribution in Sentence Amount

Prompting ChatGPT to create synthetic data and annotations

Generating Patient data using ChatGPT

Prompt:

Please write [num notes] medical notes documenting patients' [category] functioning for ICF category and level given in Context, avoid using pa- tient's age. Avoid changing patients' gender, age group, or illness. Please add a confidence score between 0 - 1 for each note. Avoid using " Patiënt" every time.

Table A.1: **Categories and Variables of "Patient"-"Medical Notes Author" Profiles**

Category	**Variables** (English/Dutch)
Illness	surgery/operatie, cancer/kanker, physiotherapy/fysiotherapie, occupational therapy/ergotherapie, dietary disorders/voedingsstoornis, chemotherapy/chemotherapie, covid
Medical Professional Title	nurse/verpleegster, physician/arts, physiotherapist/fysiotherapeut, dietitian/diëtist, occupational therapist/ergotherapeut
Gender	man/man, woman/vrouw
Age Group	18-40, 40-65, 65-80, 80+[1]

Persona

Cecilia Kuan, Text Mining Master Thesis, 2023, Vrije Universiteit Amsterdam

Synthetic data generated for ICF classification of patient records

Cecilia Kuan, Text Mining Master Thesis, 2023, Vrije Universiteit Amsterdam

Table A.6: **Example of Final Synthetic Data and "Patient"-"Medical Note Author" Profile Used - ATT Level 0**

Target:

- ICF Category and Level: b140 (ATT), 0,
- Patient: man, 18-40,
- Illness: ergotherapie / occupational therapy,
- Medical Note Author: ergotherapeut / occupational therapist.

Dutch:
output for: b140, 0, ergotherapeut, man, 18-40, ergotherapie

nootnummer |noten |confidence_score |categorie |niveau |med_personeel |geslacht |leeftijdsgroep |ziekte

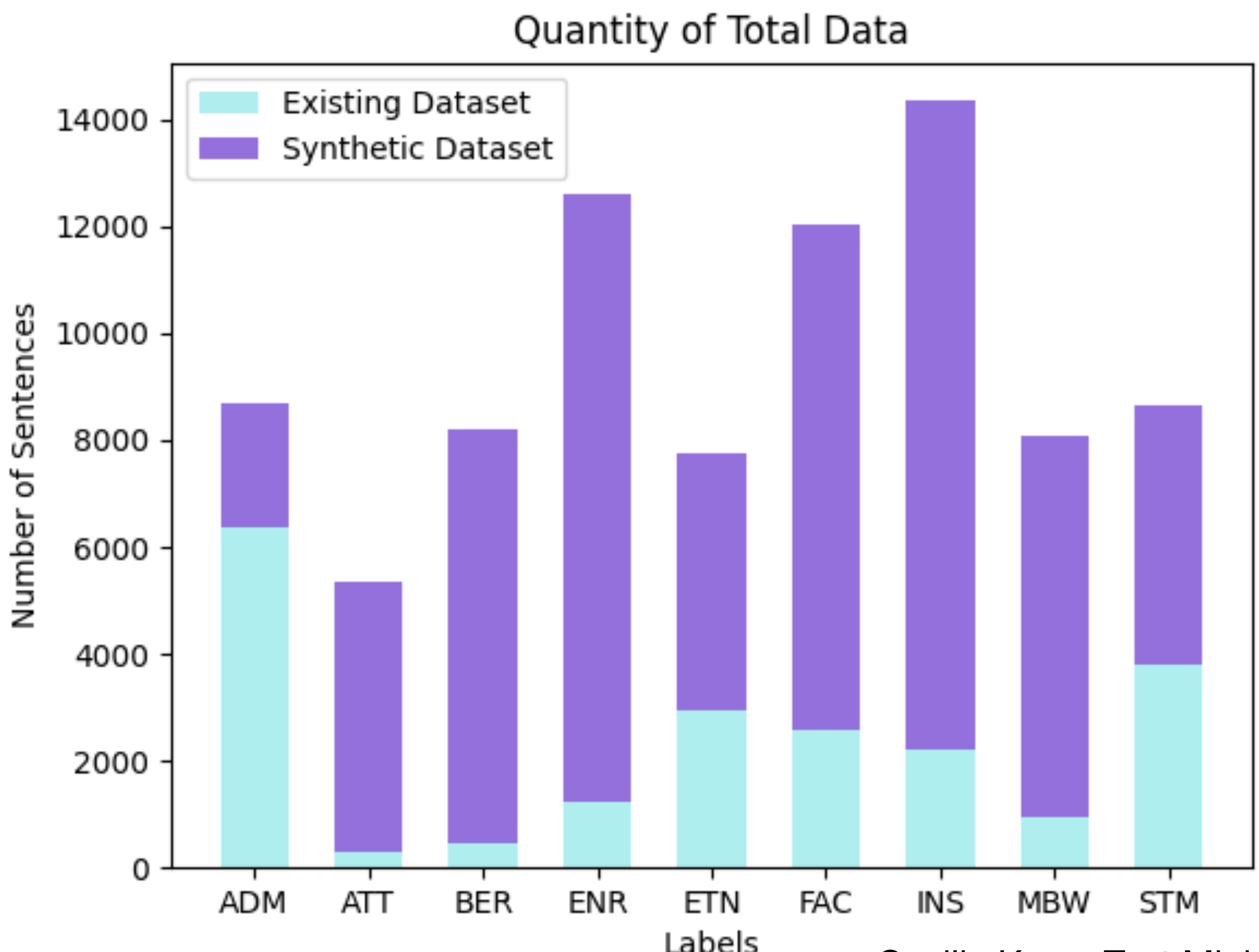
1 |Patiënt heeft moeite om zijn aandacht te concentreren op een enkele taak |0.8 |b140 |0 |ergotherapeut |man |18-40 |ergotherapie
2 |Patiënte heeft een volledige stoornis in het sturen en verdelen van haar aandacht |0.5 |b140 |0 |ergotherapeut |vrouw |18-40 |ergotherapie
3 |Patiënt kan zijn aandacht niet vasthouden gedurende een lange periode van tijd |0.9 |b140 |0 |ergotherapeut |man |18-40 |ergotherapie
4 |Patiënte heeft beperkingen in het focussen van haar aandacht |0.6 |b140 |0 |ergotherapeut |vrouw |18-40 |ergotherapie
5 |Patiënt kan zijn aandacht niet verdelen over verschillende taken |0.7 |b140 |0 |ergotherapeut |man |18-40 |ergotherapie

English:
output for: b140, 0, occupational therapist, man, 18-40, occupational therapy

note number |notes |confidence_score |category |level |med_staff|gender |age group |disease

1 |Patient has difficulty concentrating on a single task |0.8 |b140 |0 |occupational therapist |man |18-40 |occupational therapy
2 |The patient has a complete disorder in directing and dividing her attention |0.5 |b140 |0 |occupational therapist |man |18-40 |occupational therapy
3 |Patient cannot sustain attention for a long period of time |0.9 |b140 |0 |occupational therapist |man |18-40 |occupational therapy
4 |Patient has limitations in focusing her attention |0.6 |b140 |0 |occupational therapist |man |18-40 |occupational therapy
5 |Patient cannot divide his attention between different tasks |0.7 |b140 |0 |occupational therapist |man |18-40 |occupational therapy

Augmented data distribution



Cecilia Kuan, Text Mining Master Thesis, 2023, Vrije Universiteit Amsterdam

	ADM	ATT	BER	ENR	ETN	FAC	INS	MBW	STM
precision	0.98	0.98	0.56	0.96	0.92	0.84	0.89	0.79	0.70
recall	0.49	0.41	0.29	0.57	0.49	0.71	0.26	0.62	0.75
F1-score	0.66	0.58	0.35	0.72	0.63	0.76	0.41	0.70	0.72
support	775	39	54	160	382	253	287	125	181

Table 4.5: Domain classification: evaluation on test set, sentence-level

Jenia Kim, Stella Verkijk, Edwin Geleijn, Marike van der Leeden, Carel Meskers, Caroline Meskers, Sabina van der Veen, Piek Vossen, and Guy Widdershoven. Modeling Dutch Medical Texts for Detecting Functional Categories and Levels of COVID-19 Patients. Proceedings of the Language Resources and Evaluation Conference (LREC), 2022, Marseille.

-

Test: Jenia test set

	ADM	ATT	BER	ENR	ETN	FAC	INS	MBW	STM	None
precision	0.67	0.29	0.20	0.72	0.51	0.40	0.43	0.48	0.27	0.98
recall	0.61	0.51	0.78	0.59	0.80	0.83	0.57	0.81	0.74	0.92
f1-score	0.64	0.37	0.31	0.65	0.62	0.54	0.49	0.60	0.40	0.95
support	775	39	54	160	382	253	287	125	181	20013

Test: GPT dev set

	ADM	ATT	BER	ENR	ETN	FAC	INS	MBW	STM	None
precision	0.90	0.92	0.82	0.88	0.90	0.91	0.90	0.94	0.90	0.00
recall	0.81	0.91	0.70	0.81	0.93	0.90	0.87	0.87	0.94	0.00
f1-score	0.85	0.91	0.75	0.85	0.92	0.91	0.89	0.90	0.92	0.00
support	477	1050	1525	2199	946	1845	2501	1422	1000	0

Table 5.2: Classification Report of Final experiment E5 : Dual-Classifiers B2-E5, Model: MedRoBERTa.ml model

Finetuned with synthetic data generated by GPT and gold data

Cecilia Kuan, Text Mining Master Thesis, 2023, Vrije Universiteit Amsterdam

Active learning

Batch	Category	Confidence	Instances	Correct/Total	Correctness	Clustered
Batch-1	ATT	0.16-0.12	156	105/156	0.67	No
	BER	0.14-0.12	355	110/355	0.3	No
	INS	0.14-0.12	160	110/160	0.68	No
	MBW	0.14-0.12	160	106/160	0.66	No
Batch-1 Total			831	431/831	0.51	
Batch-2	ADM	0.17-0.10	83	51/83	0.61	Yes
	ATT	0.17-0.11	573	47/573	0.08	No
	BER	0.17-0.12	180	5/180	0.02	Yes
	ENR	0.16-0.12	70	62/70	0.88	Yes
	ETN	0.16-0.12	119	64/119	0.53	Yes
	FAC	0.16-0.12	62	25/62	0.40	Yes
	INS	0.14-0.12	202	39/202	0.19	Yes
	MBW	0.14-0.12	128	23/128	0.17	Yes
	STM	0.16-0.12	37	15/37	0.40	Yes
Batch-2 Total			1454	331/1454	0.22	
Batch-3	ADM	0.16-0.12	23	14/23	0.63	No
	ATT	0.16-0.12	500	5/500	0.01	No
	BER	0.14-0.12	500	6/500	0.01	No
	ENR	0.16-0.12	89	17/89	0.19	No
	ETN	0.16-0.12	112	42/112	0.375	No
	FAC	0.16-0.12	121	61/121	0.5	No
	INS	0.14-0.12	392	48/392	0.12	No
	MBW	0.14-0.12	277	10/277	0.03	No
	STM	0.16-0.12	114	65/114	0.57	No
Batch-3 Total			2128	268/2128	0.12	

Murat Ertas Kuan, Text Mining Master Thesis, 2024, Vrije Universiteit Amsterdam

	ADM	ATT	BER	ENR	ETN	FAC	INS	MBW	STM
precision	0.98	0.98	0.56	0.96	0.92	0.84	0.89	0.79	0.70
recall	0.49	0.41	0.29	0.57	0.49	0.71	0.26	0.62	0.75
F1-score	0.66	0.58	0.35	0.72	0.63	0.76	0.41	0.70	0.72
support	775	39	54	160	382	253	287	125	181

Table 4.5: Domain classification: evaluation on test set, sentence-level

Jenia Kim, Stella Verkijk, Edwin Geleijn, Marike van der Leeden, Carel Meskers, Caroline Meskers, Sabina van der Veen, Piek Vossen, and Guy Widdershoven. Modeling Dutch Medical Texts for Detecting Functional Categories and Levels of COVID-19 Patients. Proceedings of the Language Resources and Evaluation Conference (LREC), 2022, Marseille.

Category	Jenia-10			Jenia-M3			Jenia-M3.1			
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**	**Support**
ADM	0.97	0.47	0.63	0.98	0.40	0.57	0.93	**0.64**	**0.76**	891
ATT	1.00	0.41	0.58	0.95	0.49	0.64	0.70	**0.54**	**0.61**	39
BER	0.97	0.27	0.43	0.93	0.12	0.21	0.66	**0.69**	**0.67**	110
ENR	0.97	0.52	0.68	0.97	0.53	0.69	0.90	**0.65**	**0.75**	165
ETN	0.92	0.44	0.60	0.94	0.39	0.55	0.78	**0.74**	**0.76**	414
FAC	0.92	0.67	0.78	0.92	0.70	0.80	0.77	**0.79**	**0.78**	281
INS	0.77	0.16	0.27	0.73	0.27	0.40	0.54	**0.43**	**0.48**	165
MBW	0.93	0.56	0.70	0.85	0.67	0.75	0.81	**0.71**	**0.76**	147
STM	0.78	0.71	0.75	0.88	0.28	0.43	0.60	**0.83**	**0.70**	210

Murat Ertas Kuan, Text Mining Master Thesis, 2024, Vrije Universiteit Amsterdam

ICF classifier

- models: hugging face.co/cltl
- GitHub: https://github.com/cltl/a-proof-zonmw
- Docker: https://hub.docker.com/repository/docker/piekvossen/a-proof-icf-classifier

ADM=respiration
ATT=attention
BER=work
ENR=energy level
ETN=eat
FAC=walking
INS=exercise tolerance
MBW=weight
STM=mood

Patient’s Functional level

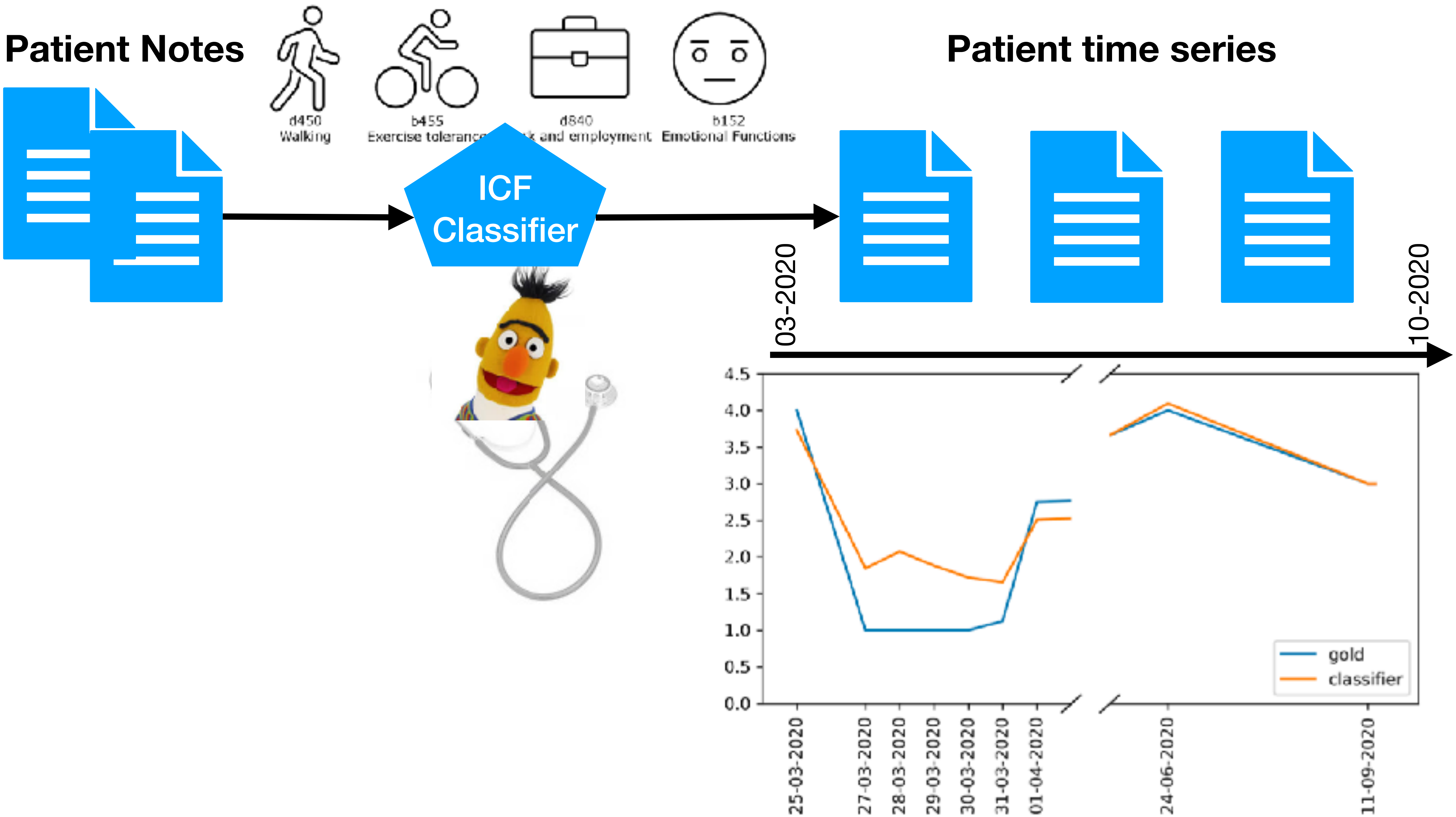


Figure 1: ADM levels of a COVID-19 patient over time

Website & Publications

- https://cltl.github.io/a-proof-project/

- Kim J, Verkijk S, Geleijn E, Van der Leeden M, Meskers CGM, Meskers CJW, Van der Veen S, Vossen PJTM, Widdershoven GAM, Modeling Dutch Medical Texts for Detecting Functional Categories and Levels of COVID-19 Patients, Proceedings of the Language Resources and Evaluation Conference (LREC), 2022, Marseille.

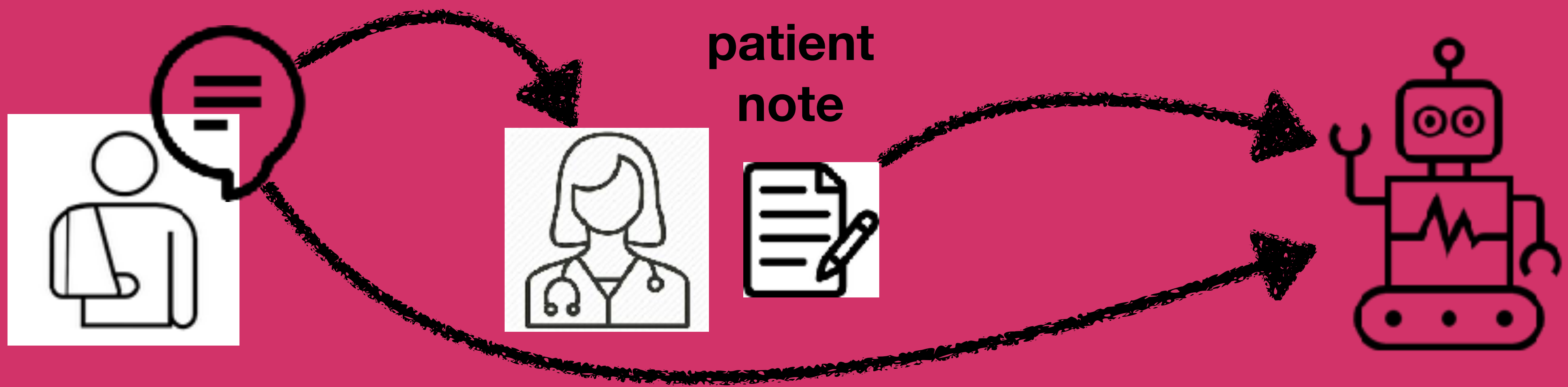
- Verkijk, S. and P. Vossen, Efficiently and thoroughly anonymizing a transformer language model for dutch electronic health records: a two-step method, 2022. Proceedings of the Language Resources and Evaluation Conference (LREC), 2022, Marseille.

- Meskers CGM, van der Veen S, Kim J, Meskers CJW, Smit QTS, Verkijk S, Geleijn E, Widdershoven GAM, Vossen PTJM, van der Leeden M. Automated recognition of functioning, activity and participation in COVID-19 from electronic patient records by natural language processing: a proof- of- concept. Ann Med. 2022 Dec;54(1):235-243

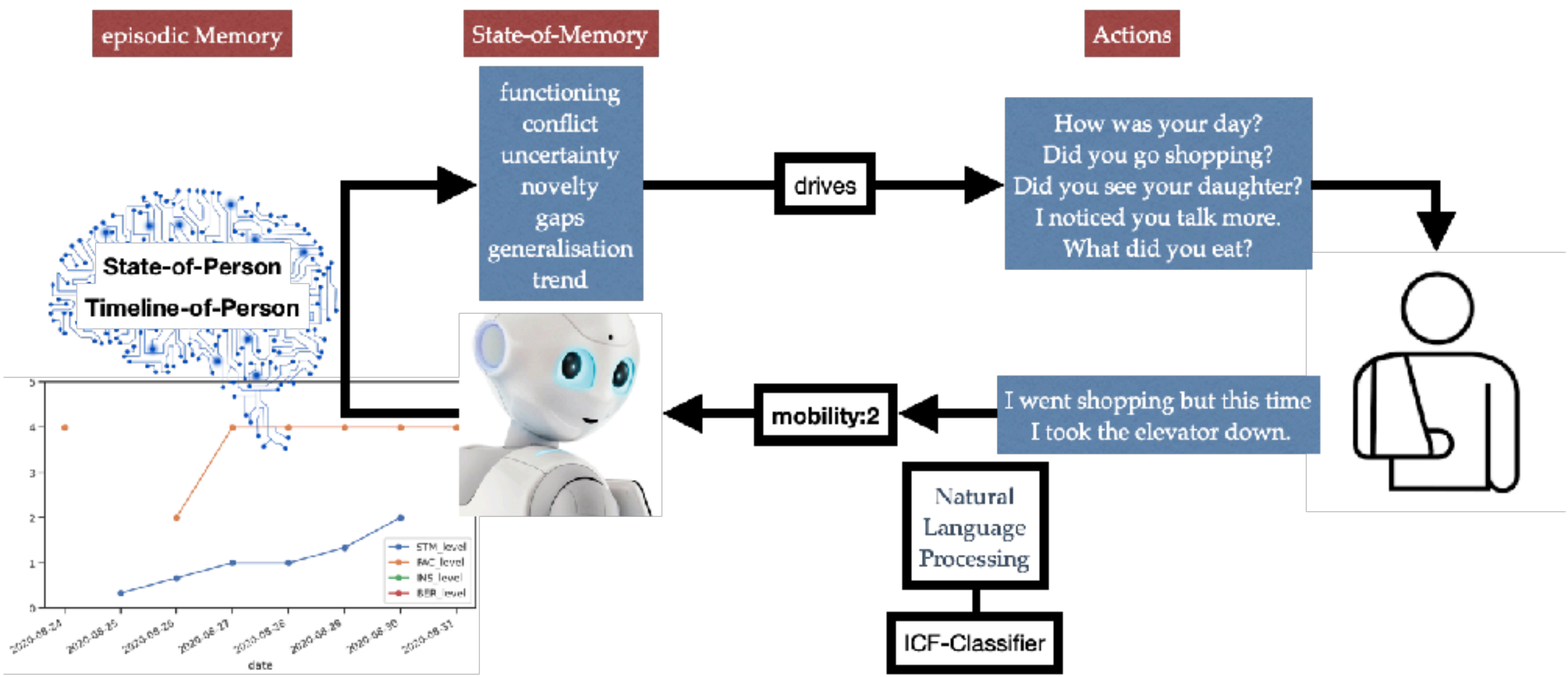
- Verkijk S, and Vossen PJTM, “Med-roBERTa.nl: a language model for dutch electronic health records,” Computational linguistics in the Netherlands journal, vol. 11, 2021

Chat4Health

Getting the patient's perspective



From Medical Notes to Conversations with patients



Structured diary of Activities of Daily Life

Piek Vossen, Selene Báez Santamaría & Thomas Baier, 2024, A Conversational Agent for Structured Diary Construction Enabling Monitoring of Functioning & Well-being, Proceedings of HHAI-2024, Sweden

Structured diary as an episodic Knowledge Graph (eKG) of events

Carl

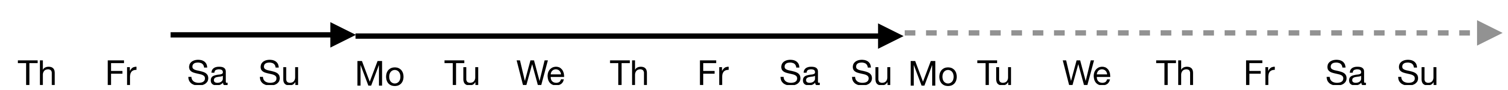
Timeline

Th Fr Sa Su Mo Tu We Th Fr Sa Su Mo Tu We Th Fr Sa Su

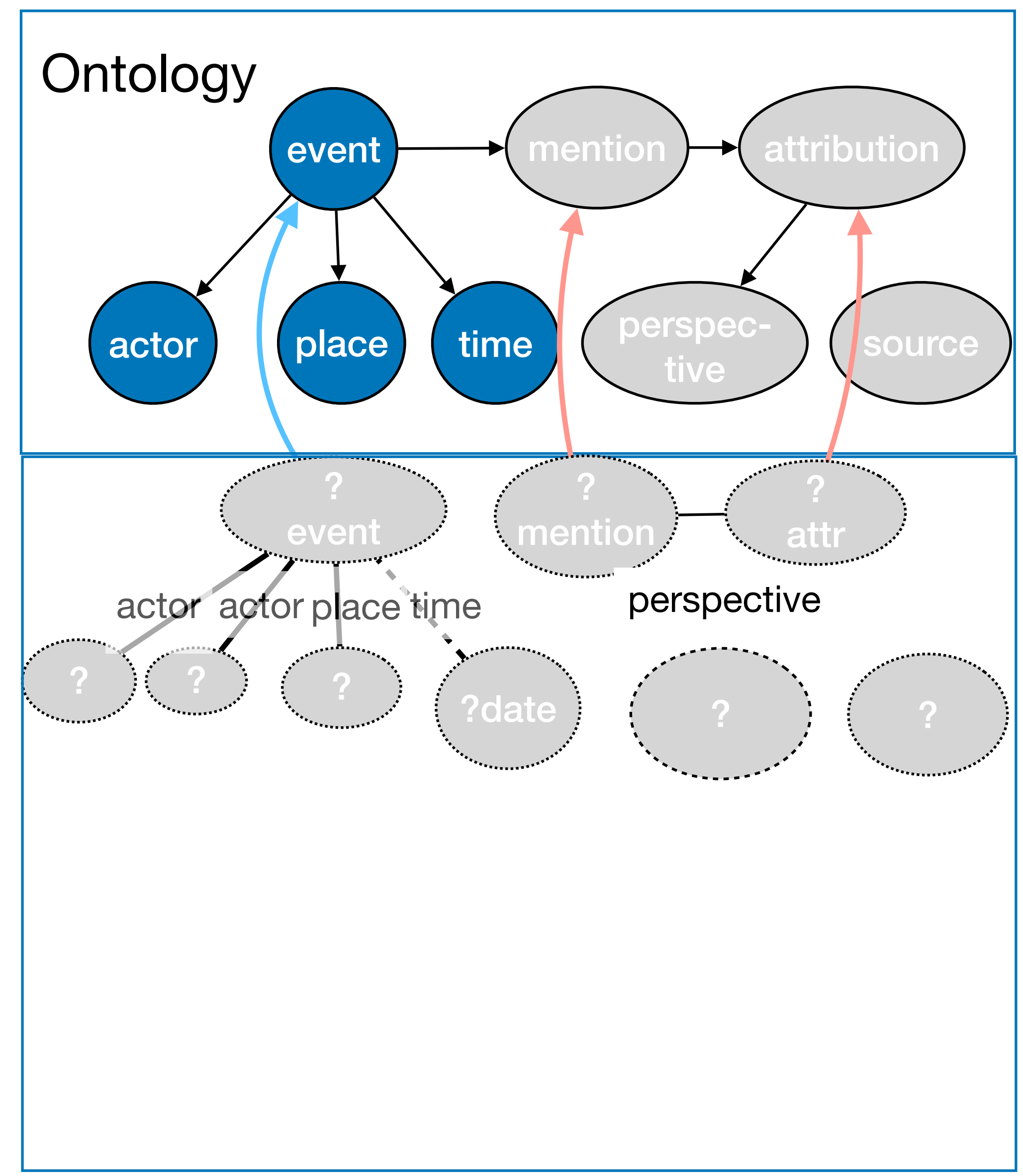
Structured diary as an episodic Knowledge Graph (eKG) of events

Carl

Timeline



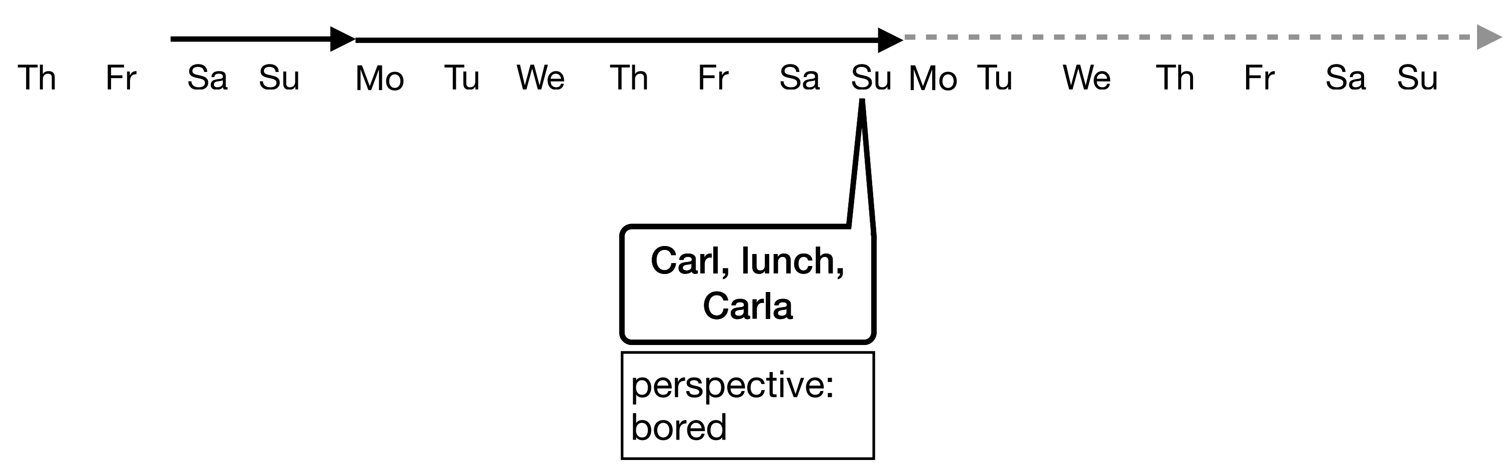
eKG



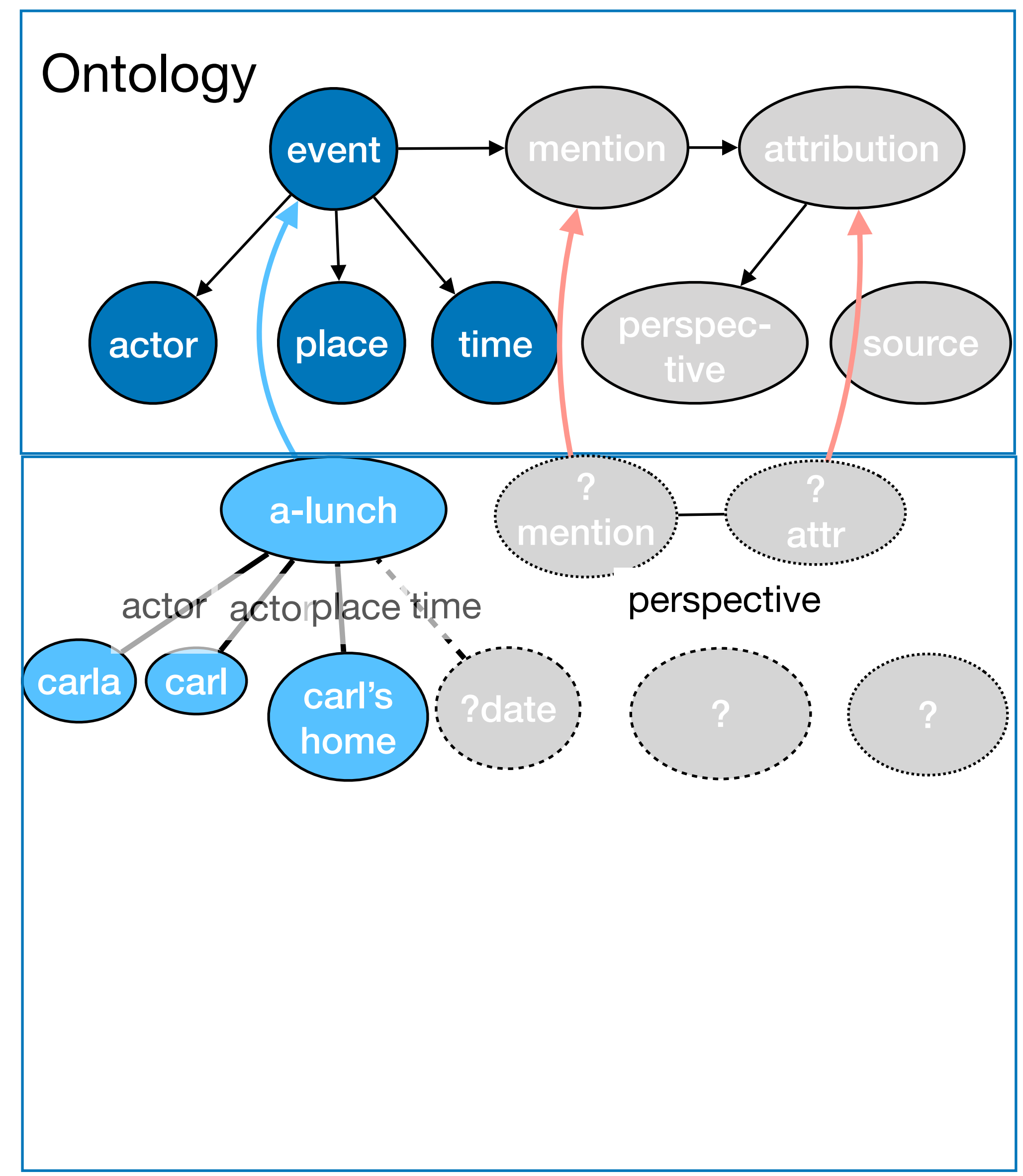
Structured diary as an episodic Knowledge Graph (eKG) of events

Carl

Timeline



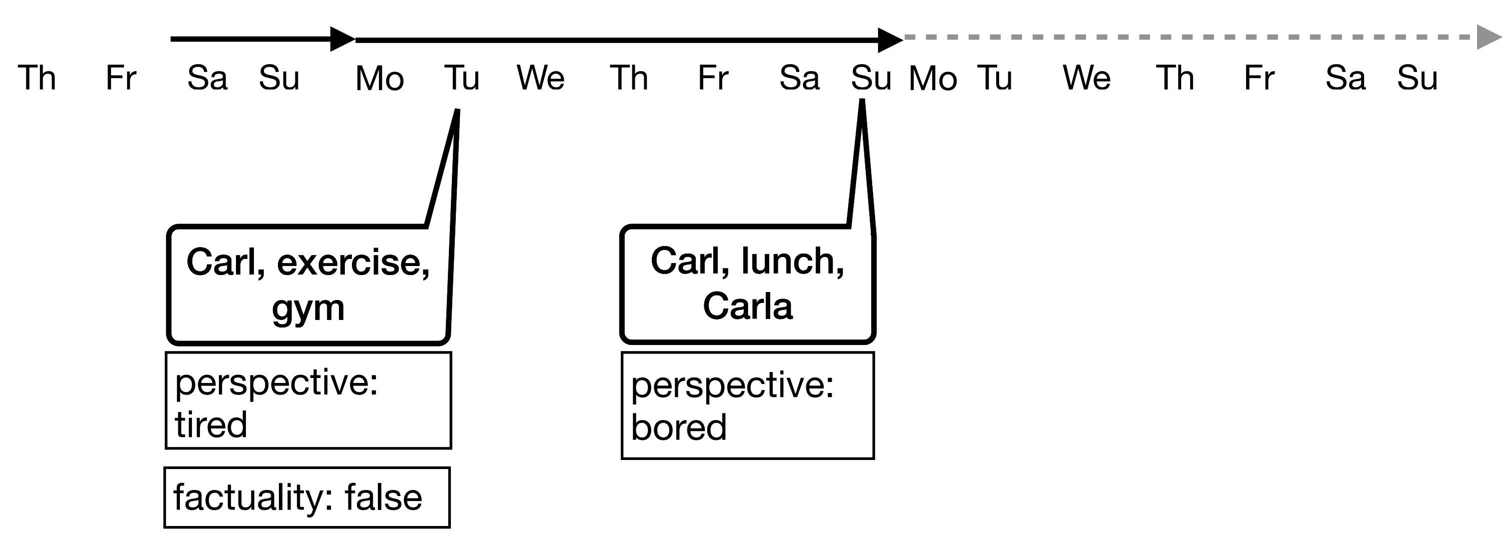
eKG



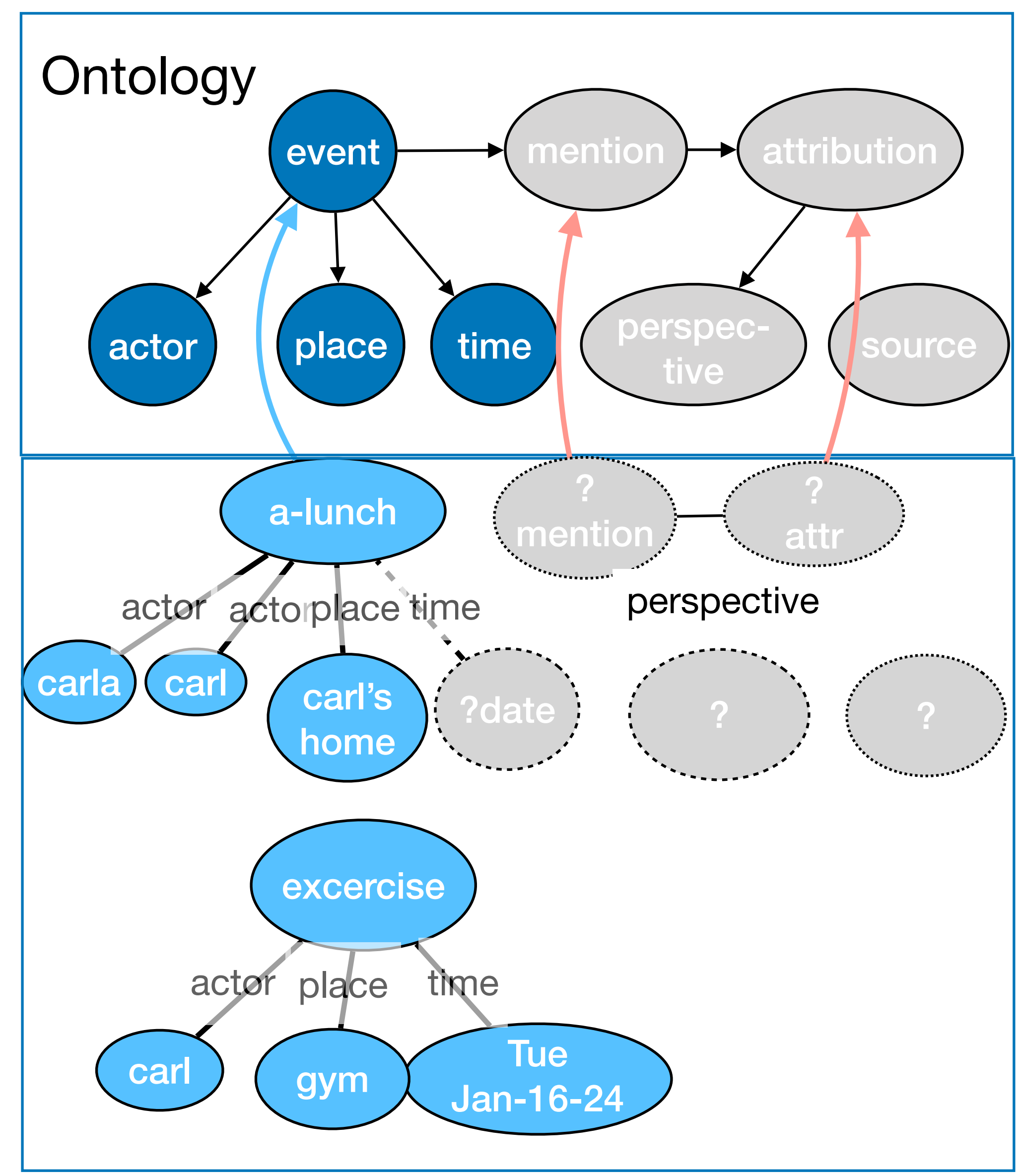
Structured diary as an episodic Knowledge Graph (eKG) of events

Carl

Timeline



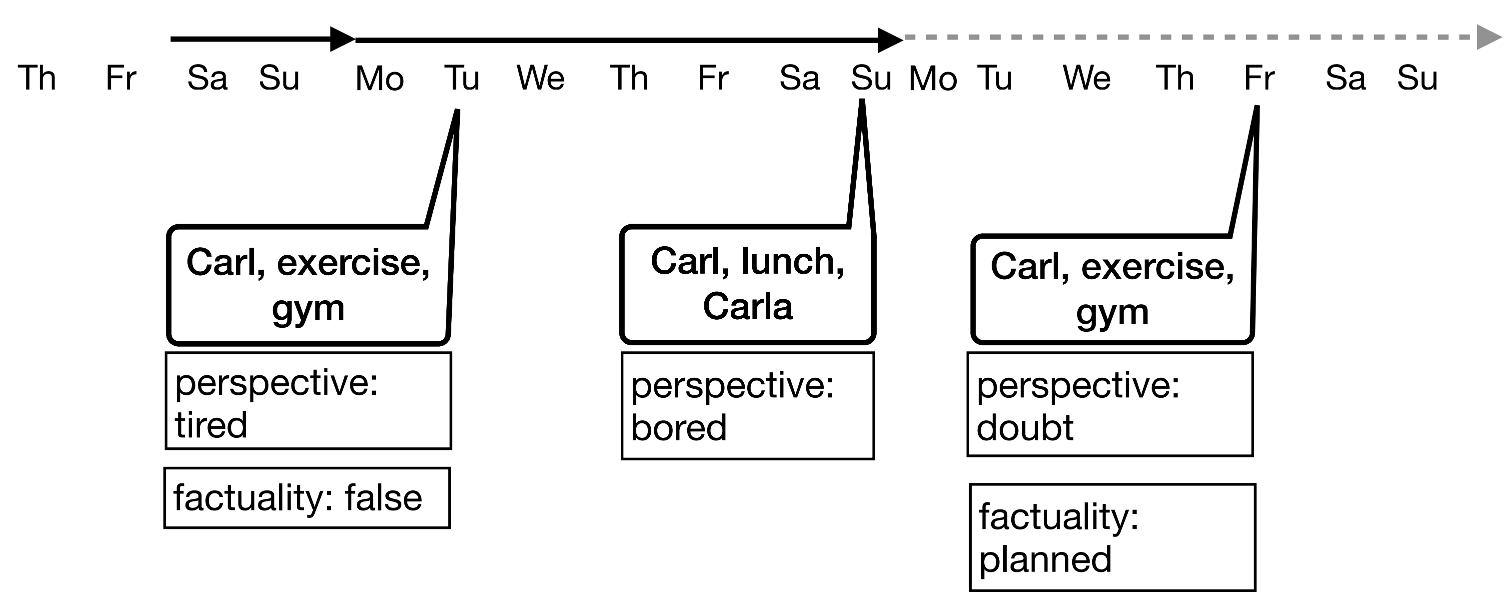
eKG



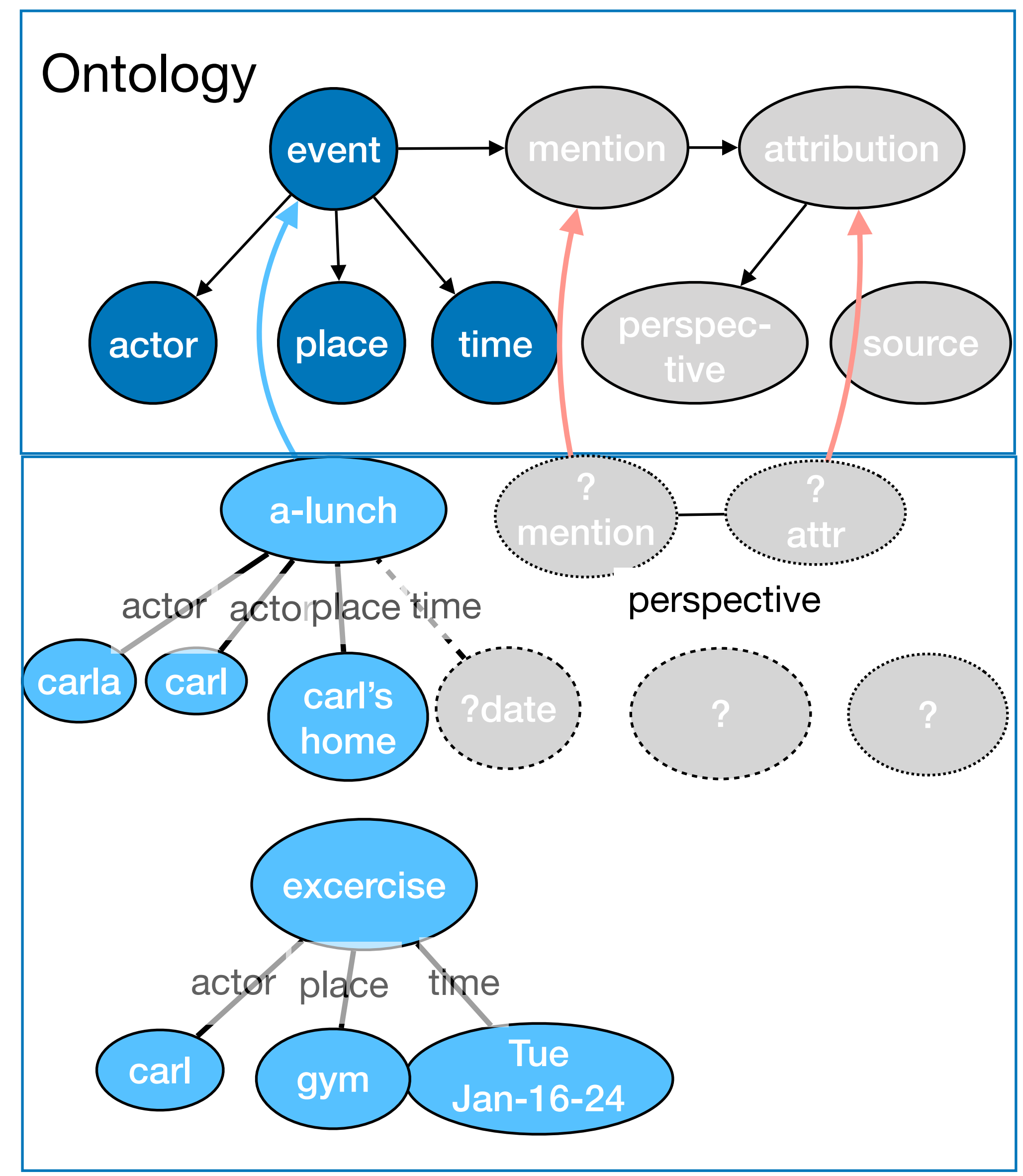
Structured diary as eKG of events

Carl

Timeline



eKG

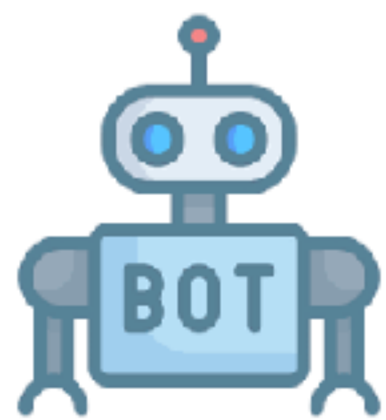


Conversational Agent

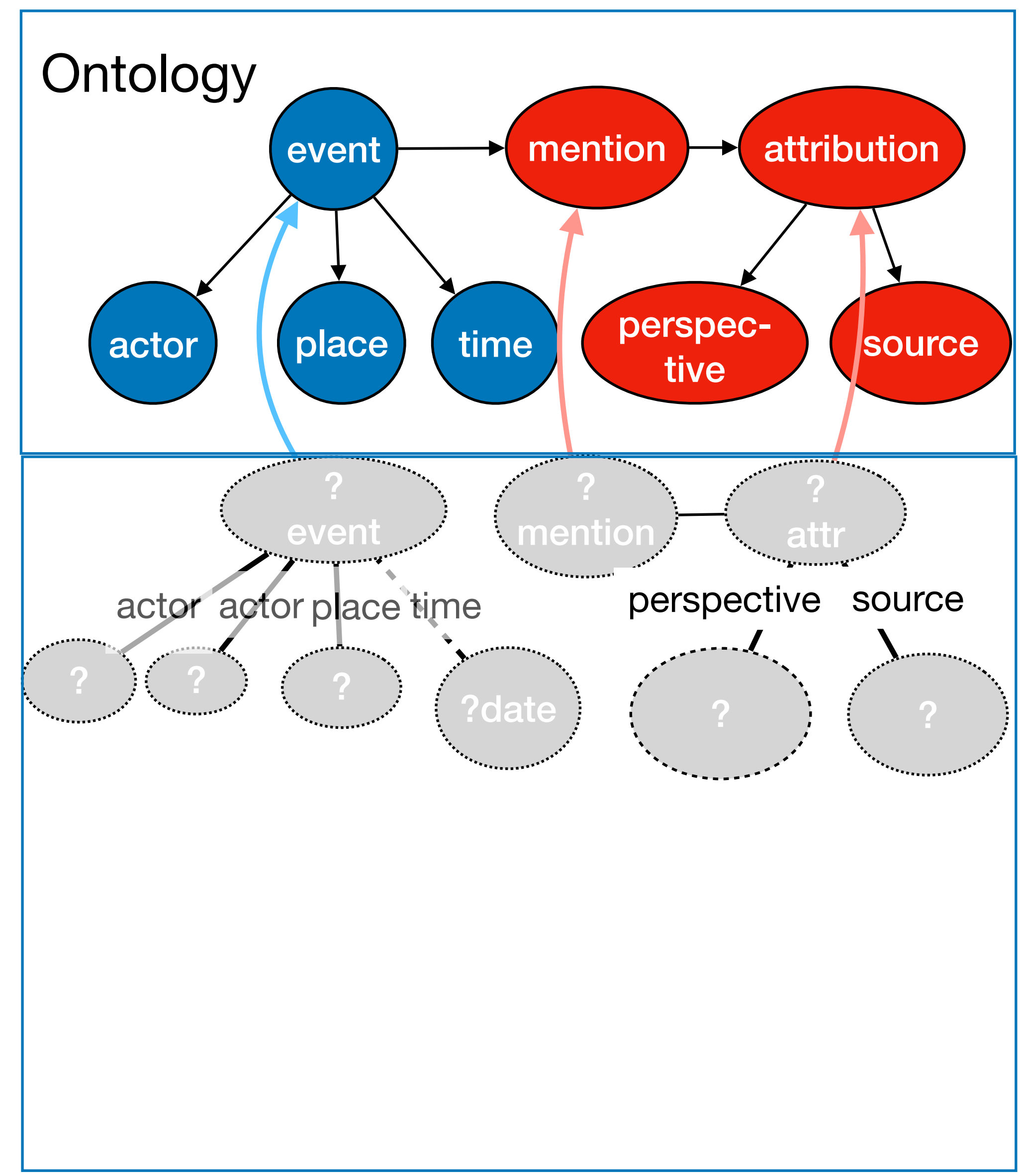
Piek Vossen, Selene Báez Santamaría & Thomas Baier, 2024, A Conversational Agent for Structured Diary Construction Enabling Monitoring of Functioning & Well-being, Proceedings of HHAI-2024, Sweden

Collecting data through conversation

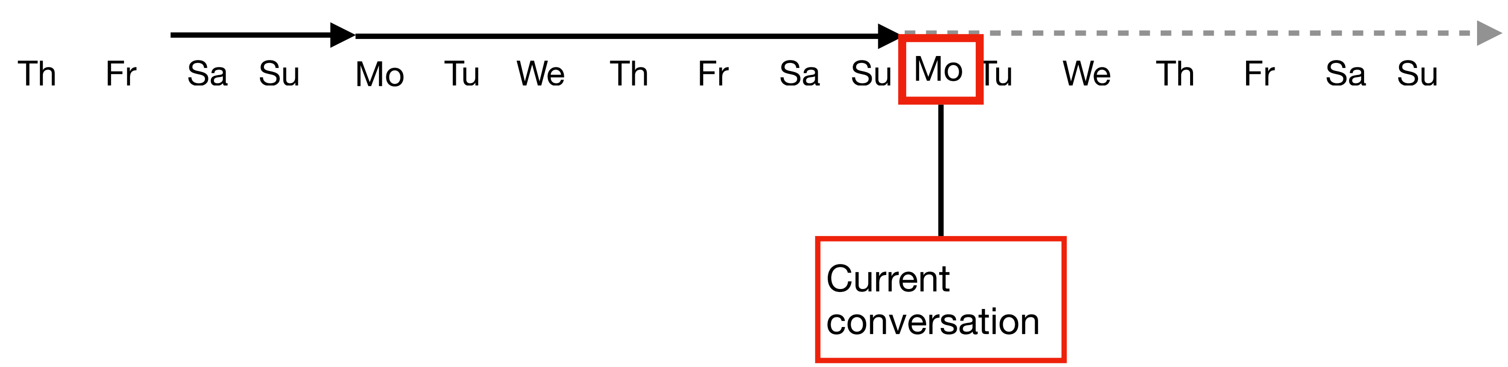
Carl



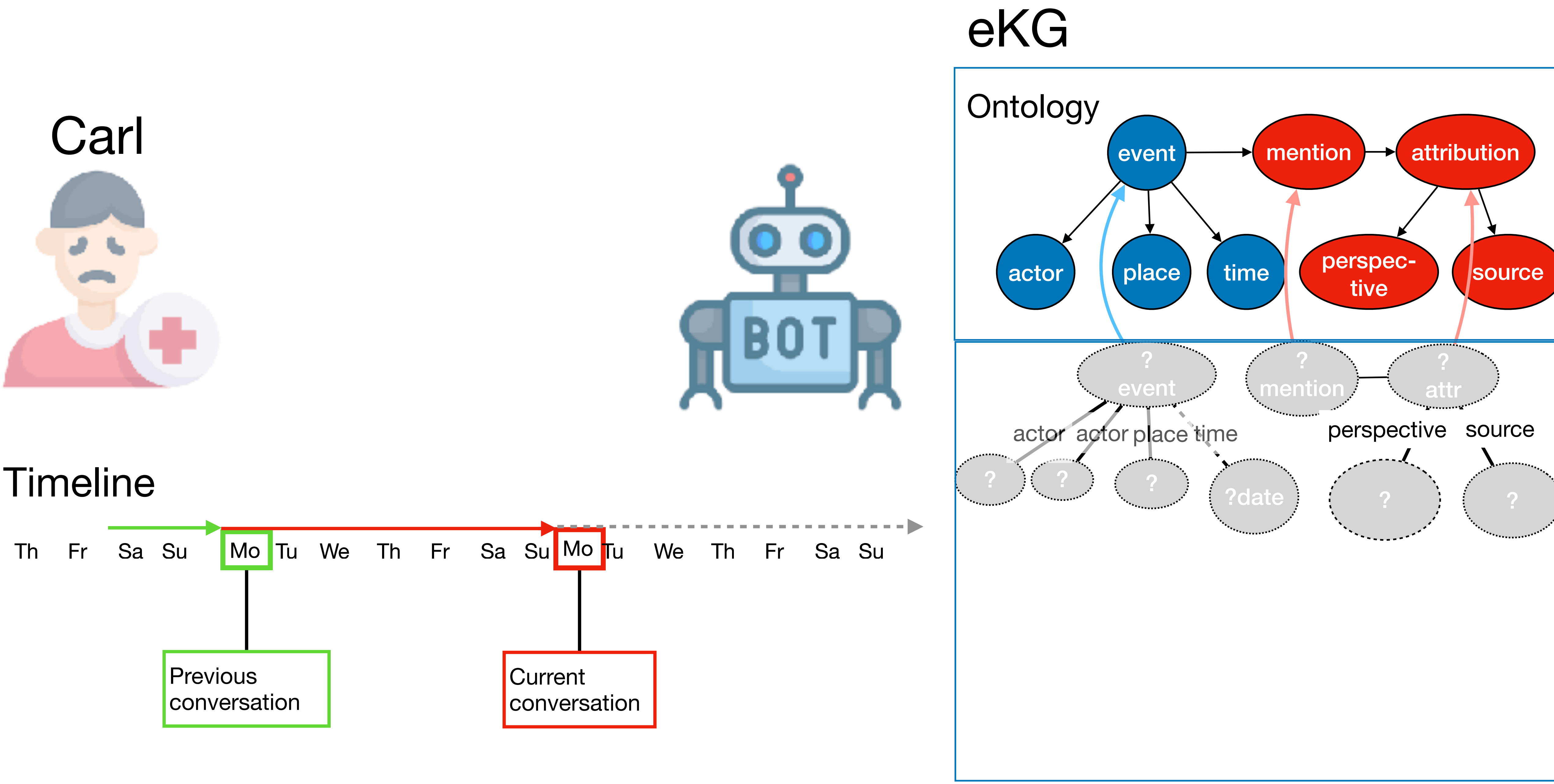
eKG



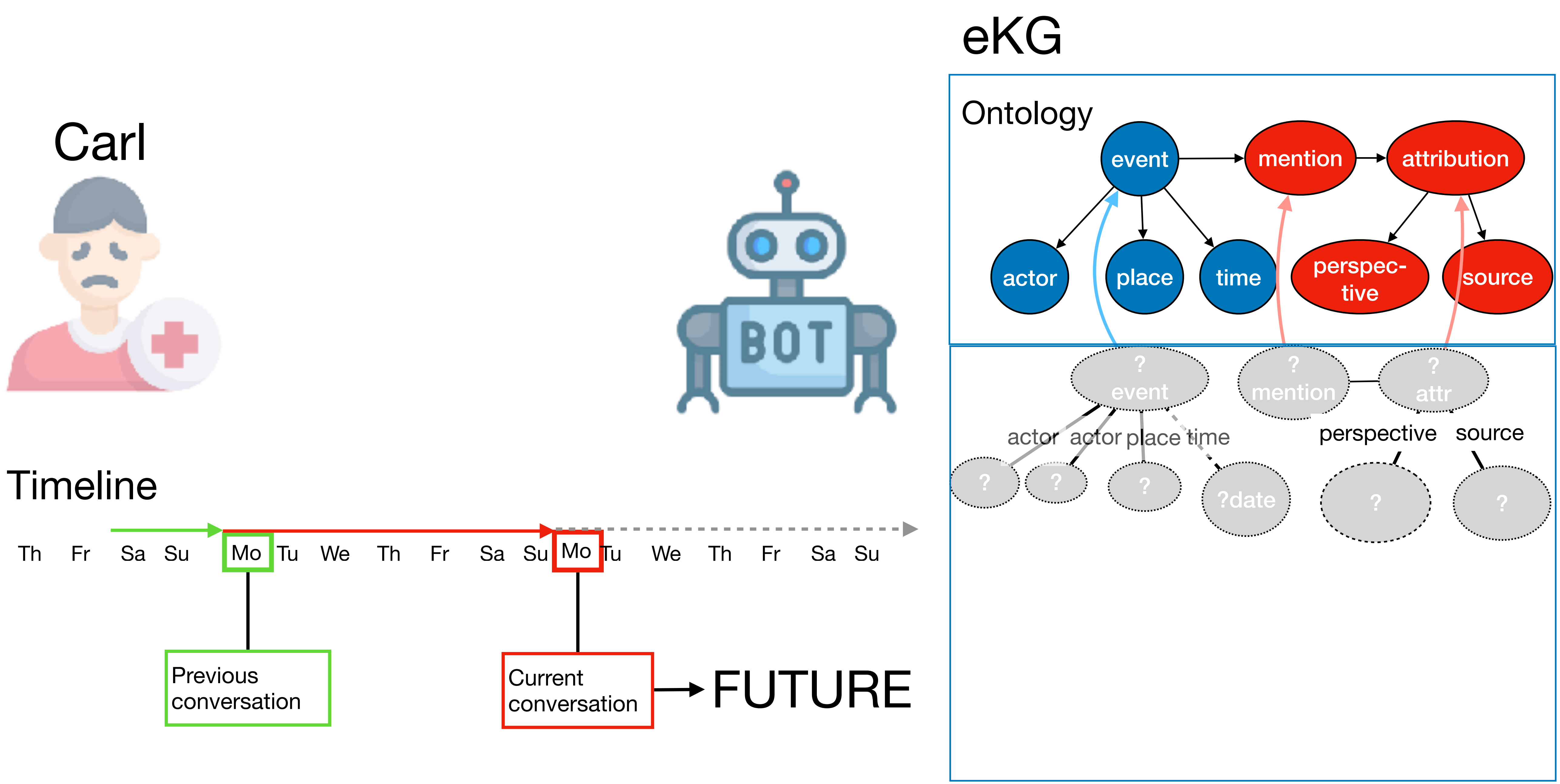
Timeline



Collecting data through conversation

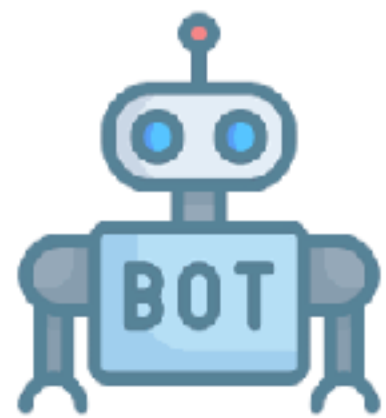


Collecting data through conversation

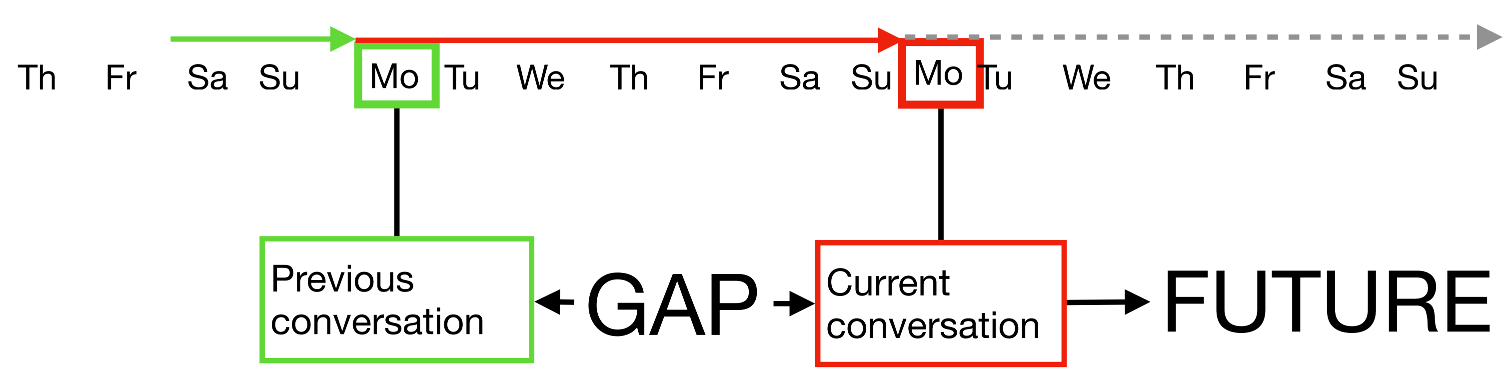


Collecting data through conversation

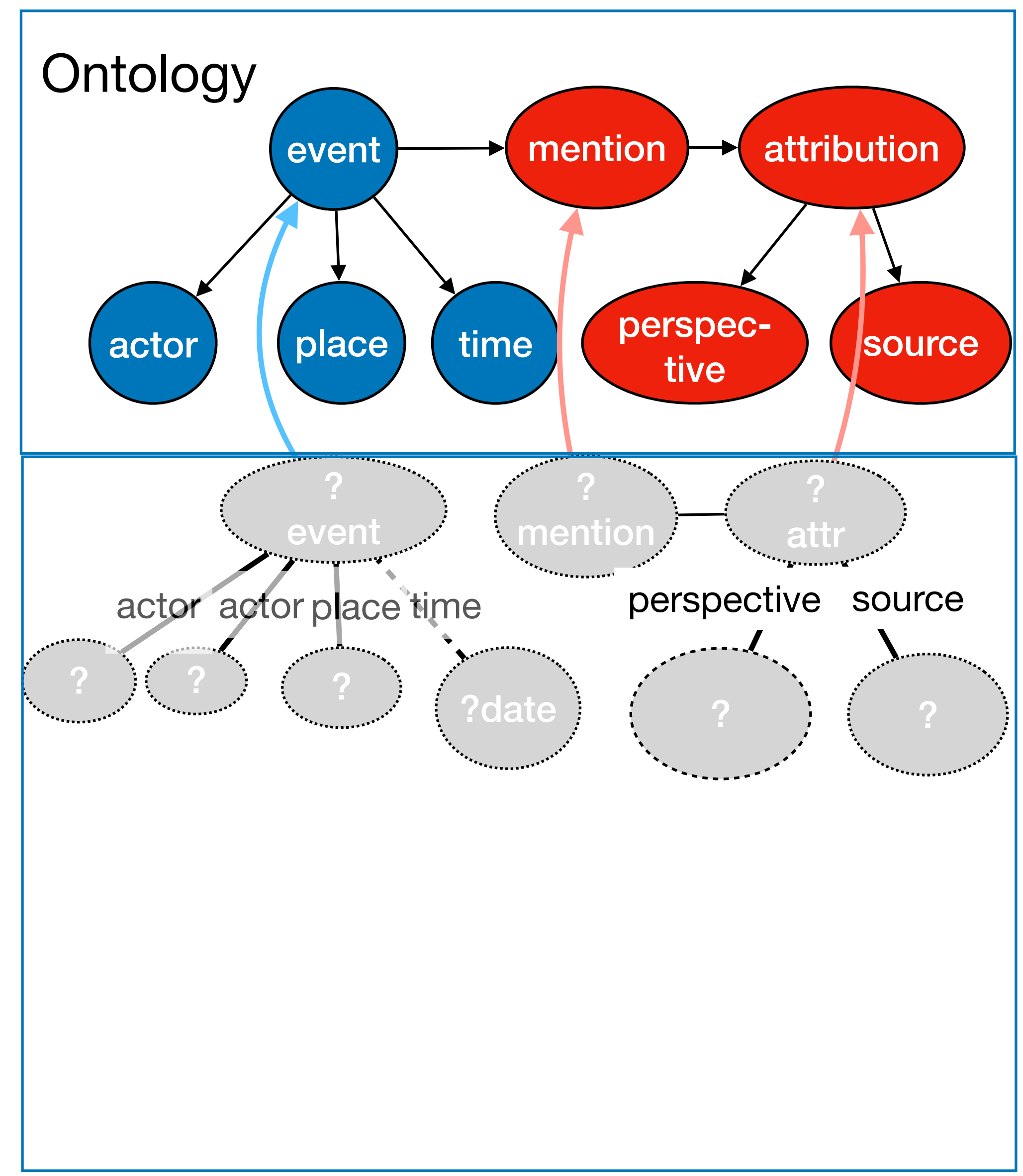
Carl



Timeline



eKG



Collecting data through conversation

Drives of the conversation (Desires & Intents):

1. High level desires driven by the timeline:
    1. Did what was expected/planned happen: Did you had that lunch with your sister?
    2. What happened since the last conversation: How was your weekend?
    3. What is planned to happen: What are your plans for tomorrow?
2. Mid level intents driven by necessary/possible/probable properties in the ontology:
    1. Who, what, when, where: What did you had for lunch? Analogies: Did she stay long again?
    2. Probabilities: Did you drink wine?
3. Low level intents driven by possible and likely perspectives:
    1. emotion: How was it for you? certainty: Are you sure?
    2. conflicts: But you told me before that...

Collecting data through conversation

Drives of the conversation (Desires & Intents):

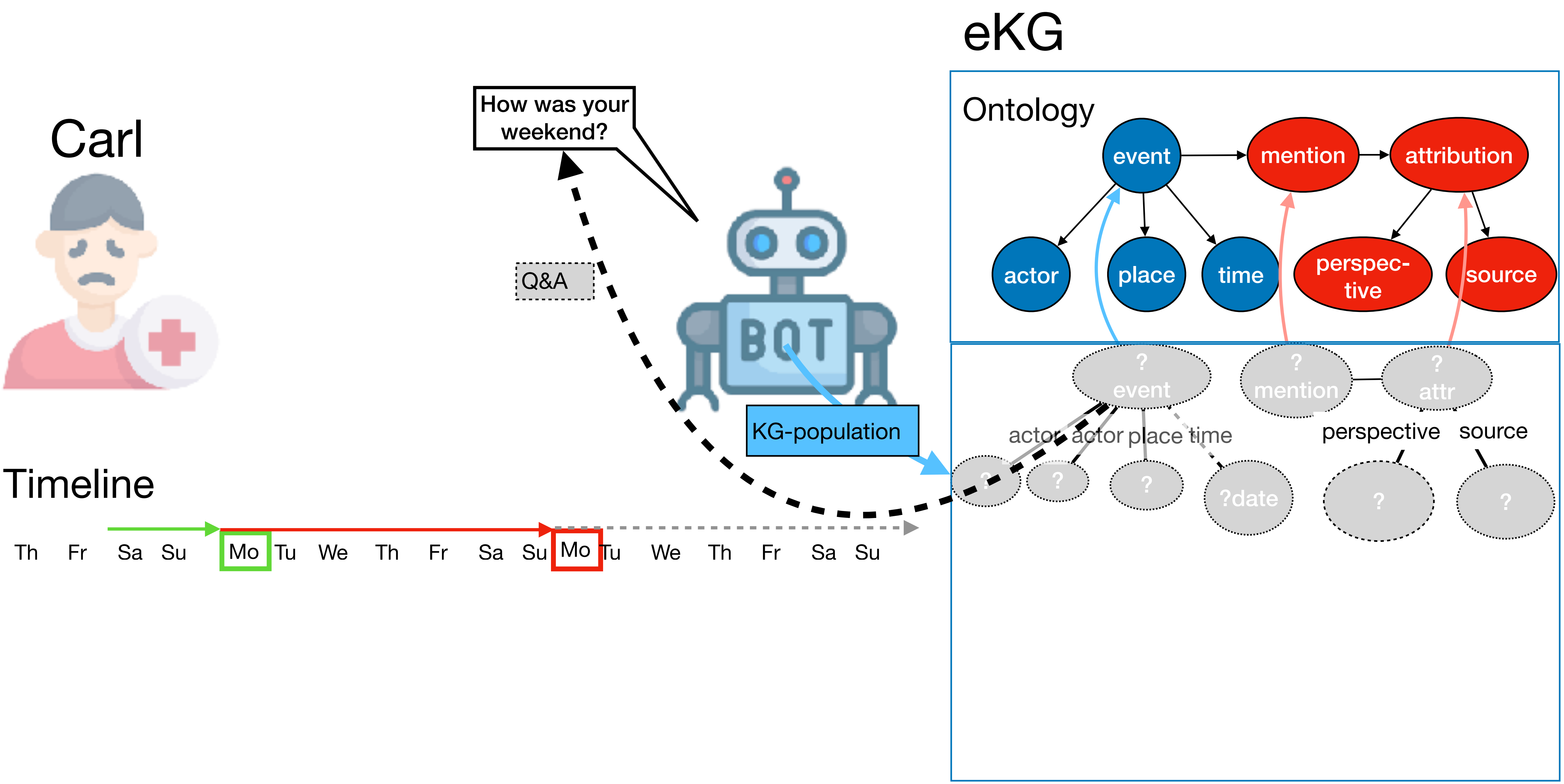
1. High level desires driven by the timeline:
    1. Did what was expected/planned happen: Did you had that lunch with your sister?
    2. What happened since the last conversation: How was your weekend?
    3. What is planned to happen: What are your plans for tomorrow?
2. Mid level intents driven by necessary/possible/probable properties in the ontology:
    1. Who, what, when, where: What did you had for lunch? Analogies: Did she stay long again?
    2. Probabilities: Did you drink wine?
3. Low level intents driven by possible and likely perspectives:
    1. emotion: How was it for you? certainty: Are you sure?
    2. conflicts: But you told me before that...

Collecting data through conversation

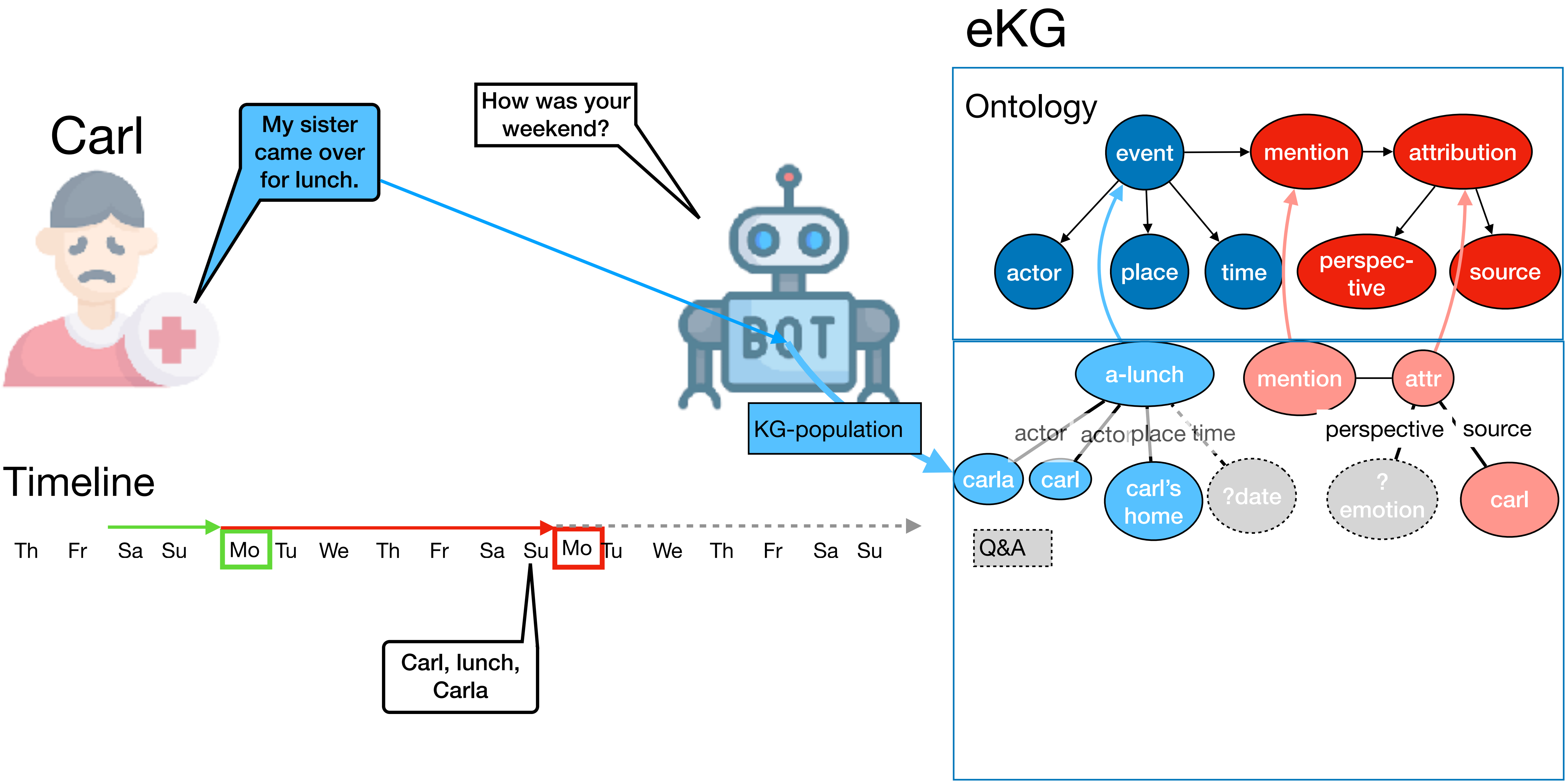
Drives of the conversation (Desires & Intents):

1. High level desires driven by the timeline:
    1. Did what was expected/planned happen: Did you had that lunch with your sister?
    2. What happened since the last conversation: How was your weekend?
    3. What is planned to happen: What are your plans for tomorrow?
2. Mid level intents driven by necessary/possible/probable properties in the ontology:
    1. Who, what, when, where: What did you had for lunch? Analogies: Did she stay long again?
    2. Probabilities: Did you drink wine?
3. Low level intents driven by possible and likely perspectives:
    1. emotion: How was it for you? certainty: Are you sure?
    2. conflicts: But you told me before that...

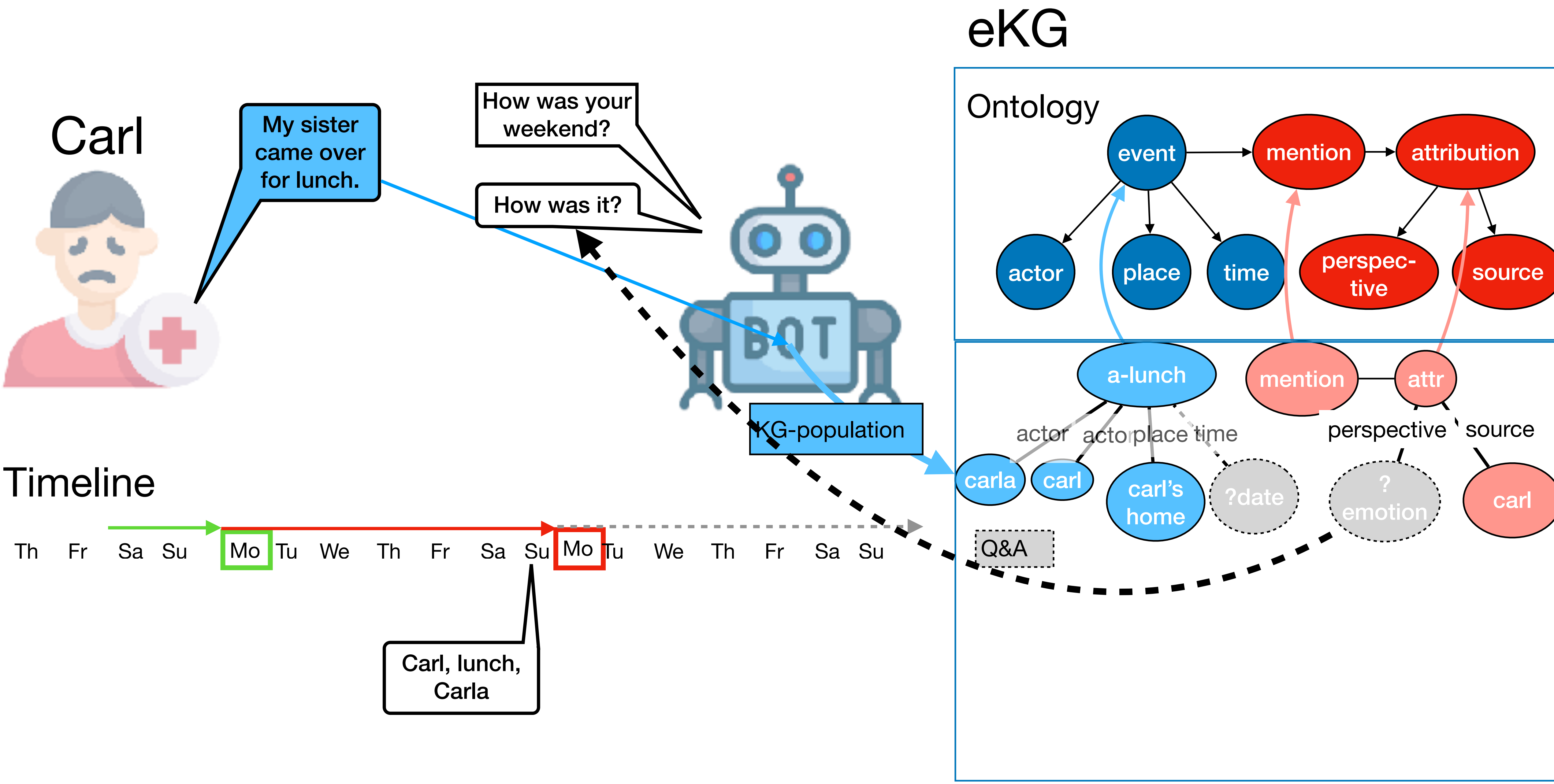
Collecting data through conversation



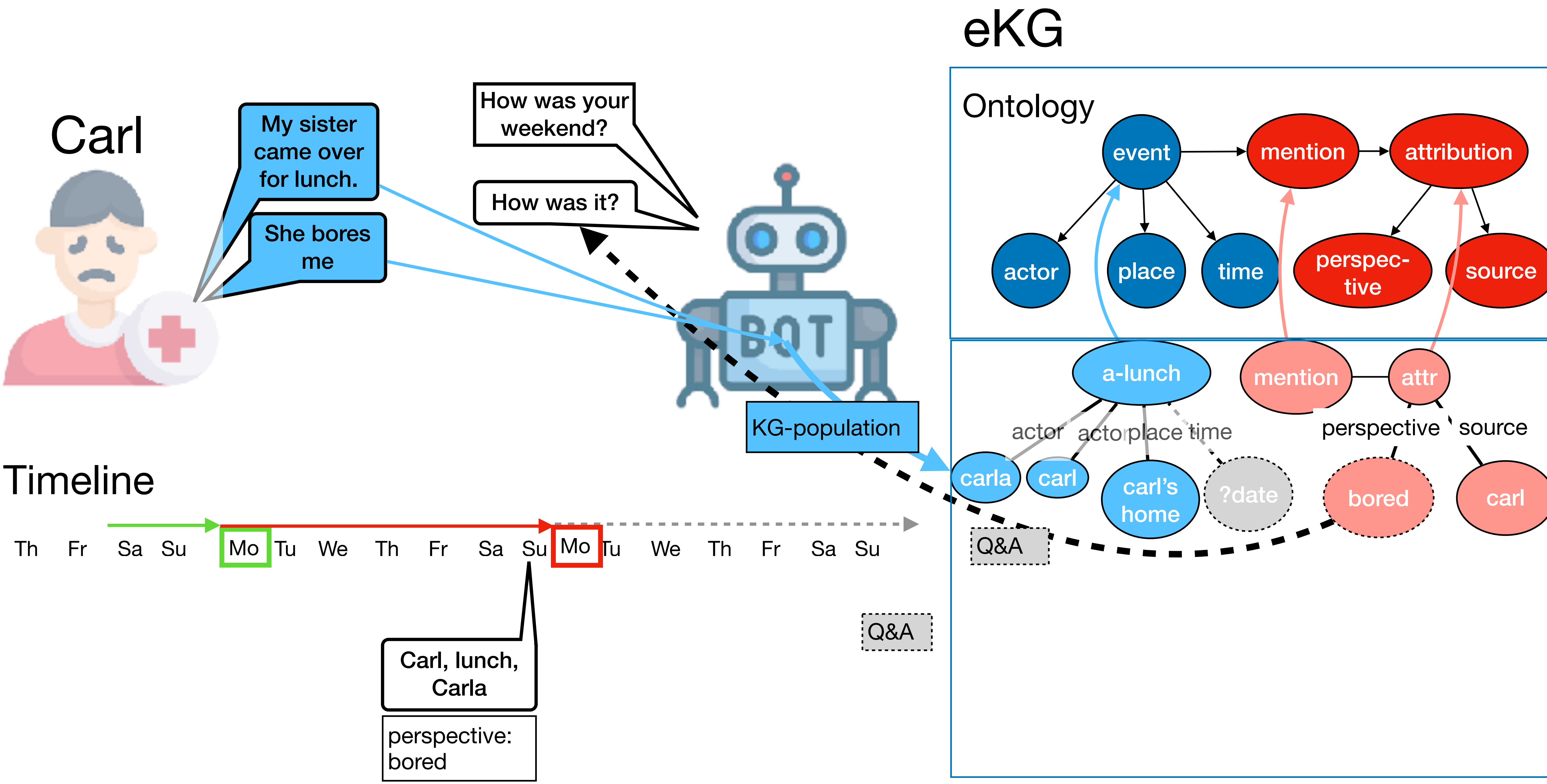
Collecting data through conversation



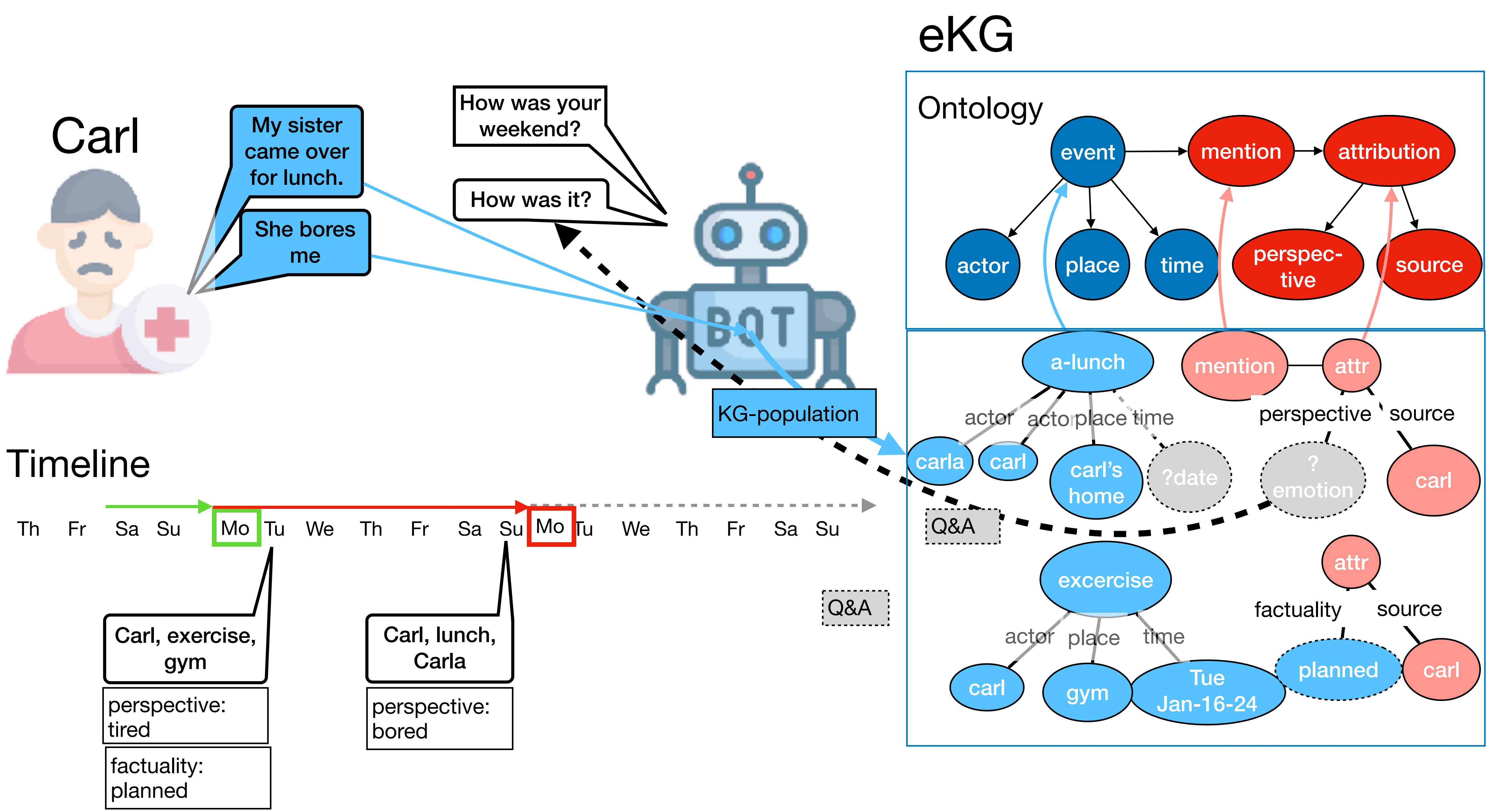
Collecting data through conversation



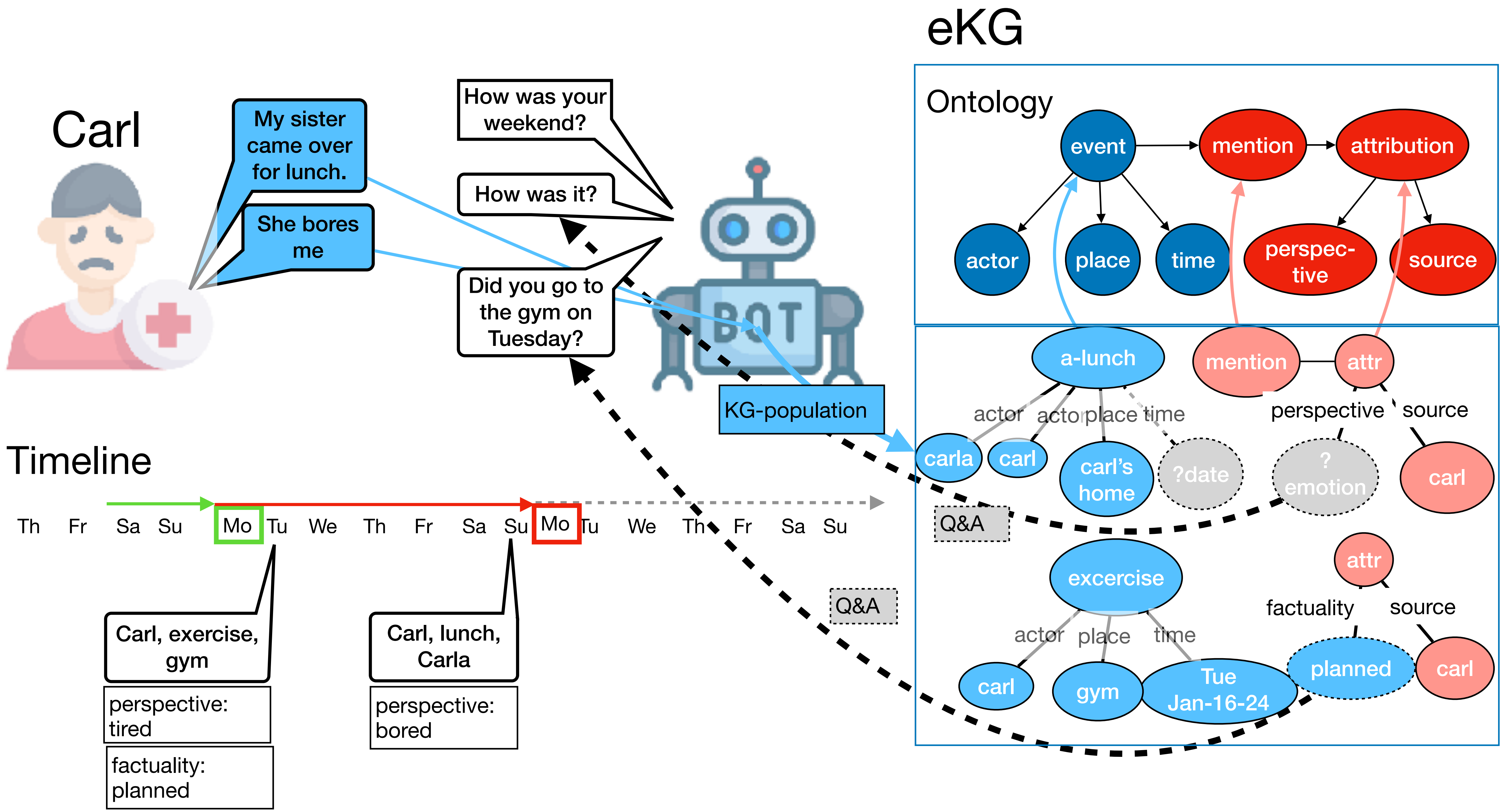
Collecting data through conversation



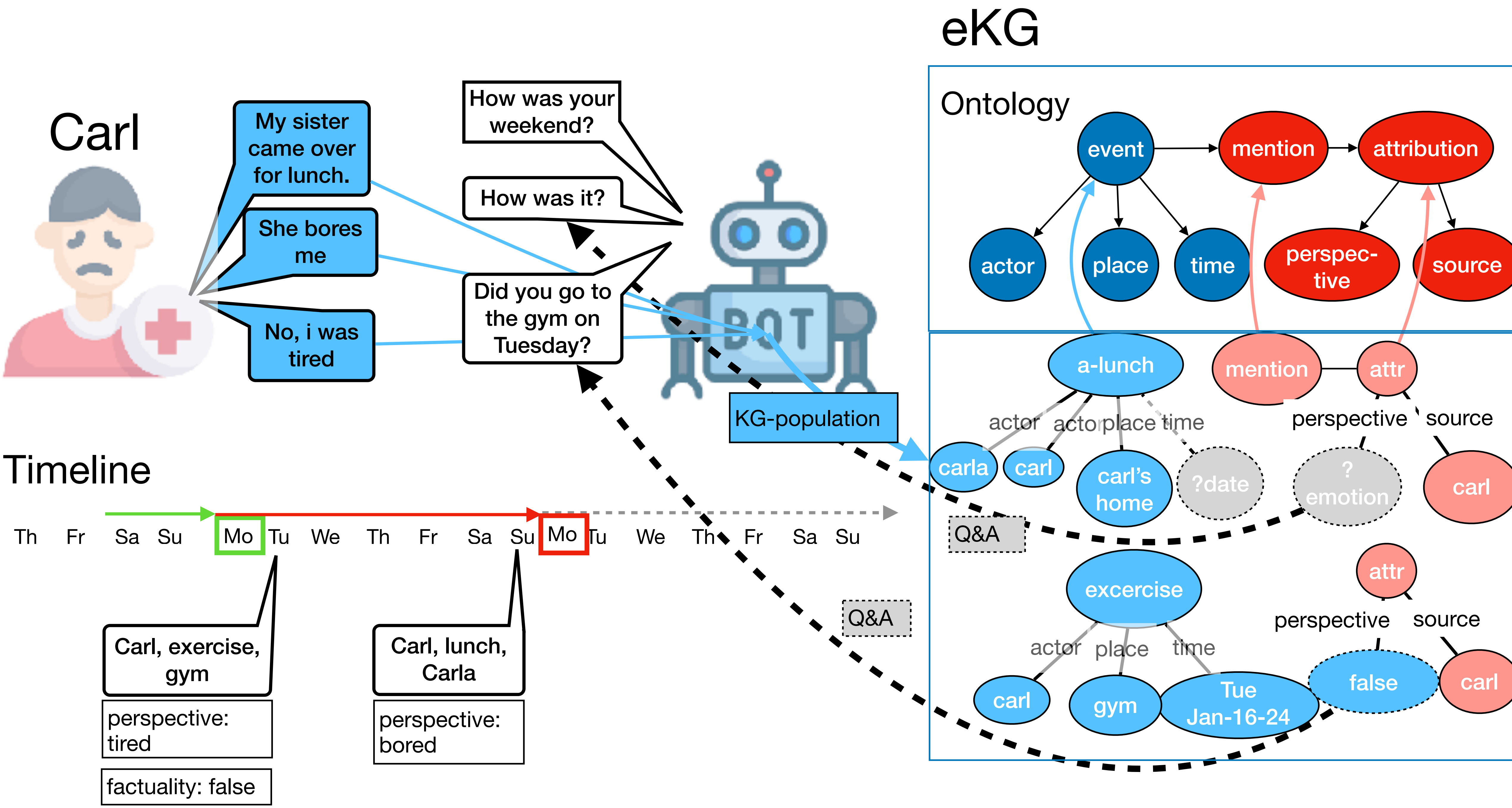
Collecting data through conversation



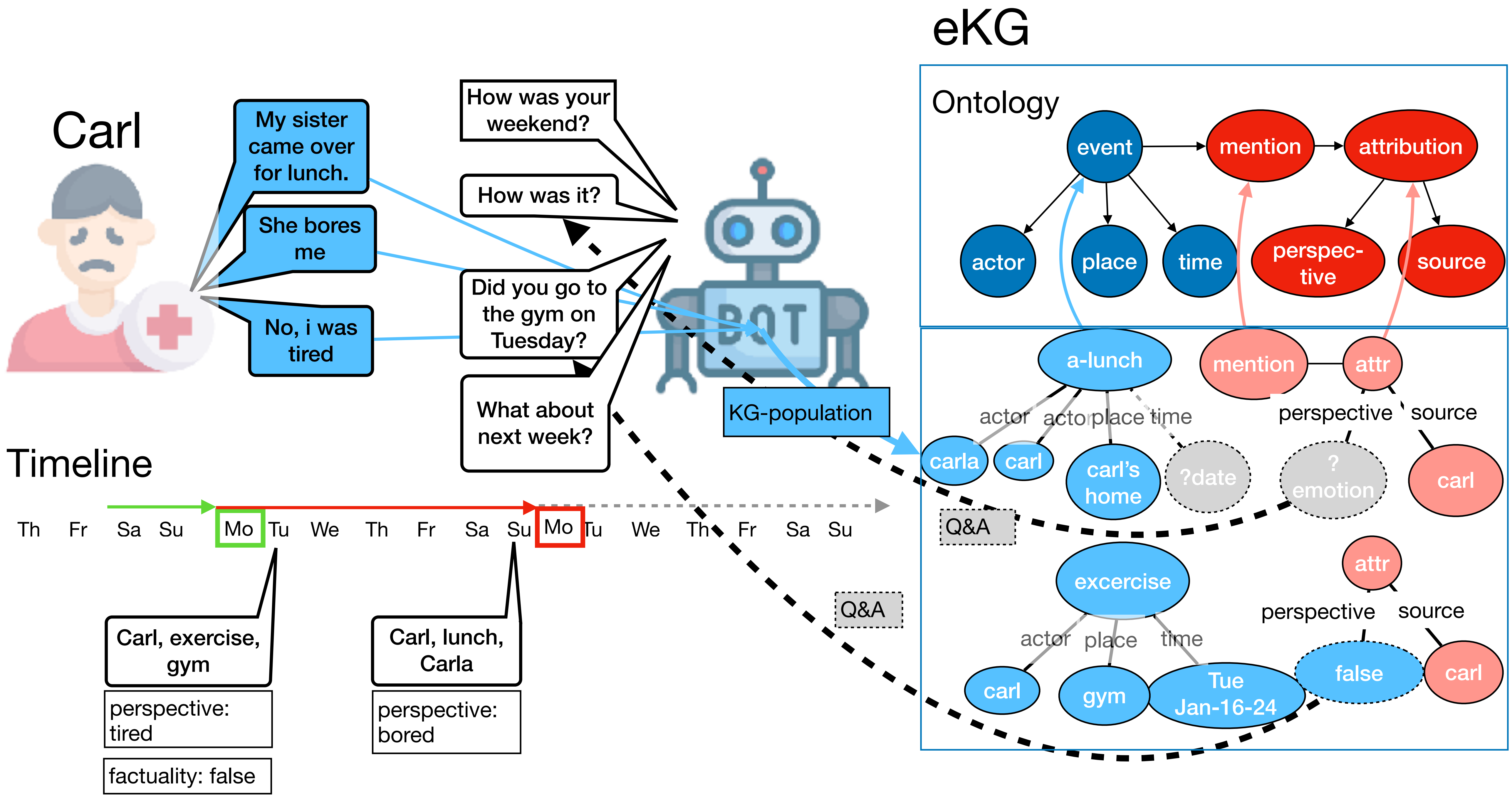
Collecting data through conversation



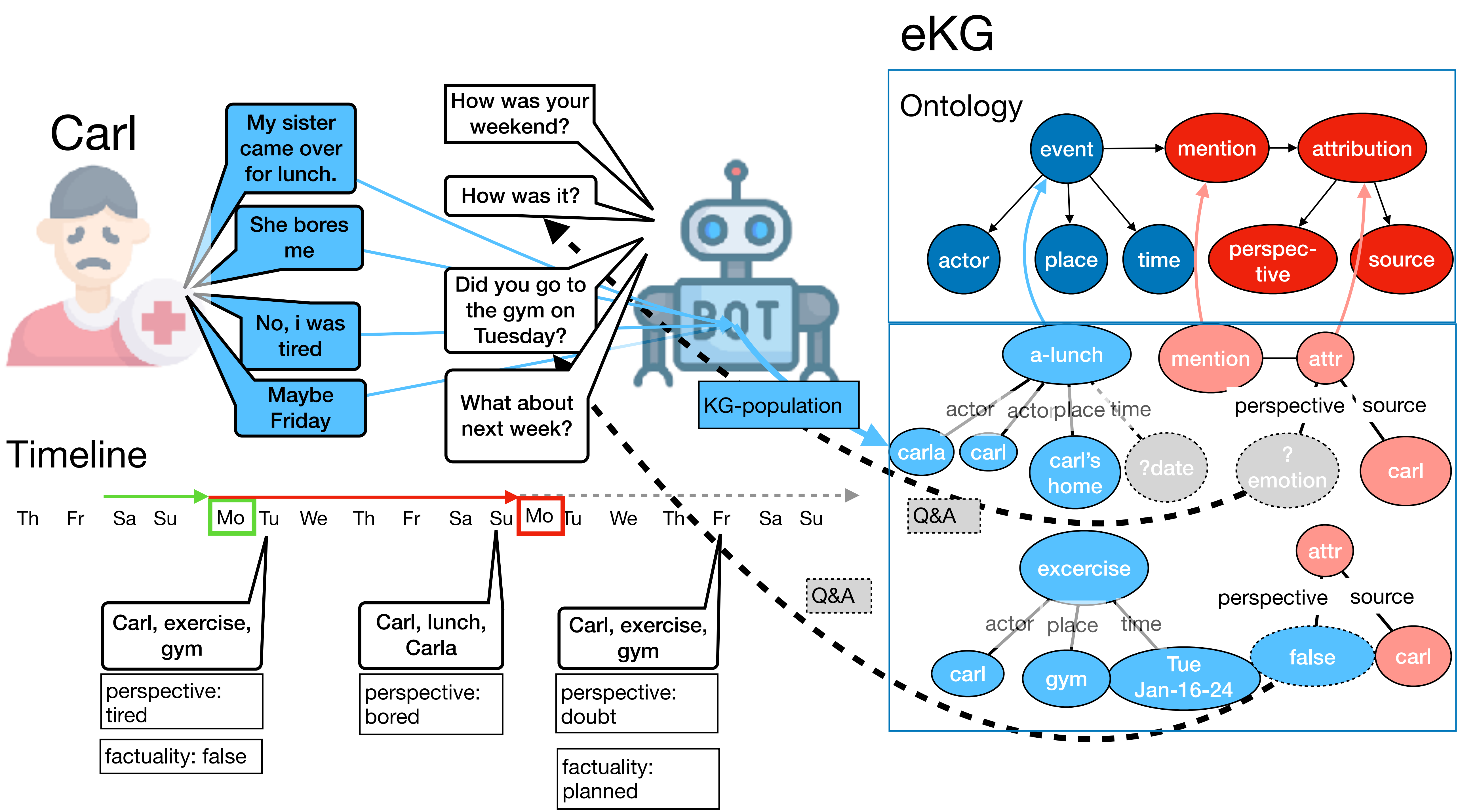
Collecting data through conversation



Collecting data through conversation



Collecting data through conversation



How to build a conversational AI for the medical domain?

- Create synthetic data with Generative AI (LLM):
  - Conversations between patient and caretaker
  - Label conversations for activity information using few shot prompting
- Finetune a BERT model to extract the information from the conversation using the synthetic data
- Finetune a generative AI (e.g. Llama3-7B) with the synthetic conversations to ask the right questions

From ICF topics to activities

- ICF categories are not sufficiently precise to detect activities: *Exercise Tolerance* — versus — *I took the stairs yesterday*
- Detect mentions of specific activities in conversations and their properties: *who* did *what*, *when* and *where, how*
- Token classification task to find phrases that represent event information

Today: Monday, June 24th, 2024

Chatbot: How was your day today [B-Time] ?

Patient: I [B-Actor] spent the morning [B-Time] cleaning [B-Activity] the [I-Activity] house [I-Activity] [B-Location .

Chatbot: How did it go?

Patient: I [B-How] had [I-How] to [I-How] sit [I-How] down [I-How] several [I-How] times [I-How] .

==> **Activity**: Clean-the-house, **Participant**: Patient, **When**: June 24th, 2024, **Where**: Home, **How**: sit-down-several-times

Chuqiao Guo, Text Mining Master Thesis 2024

Which activities?

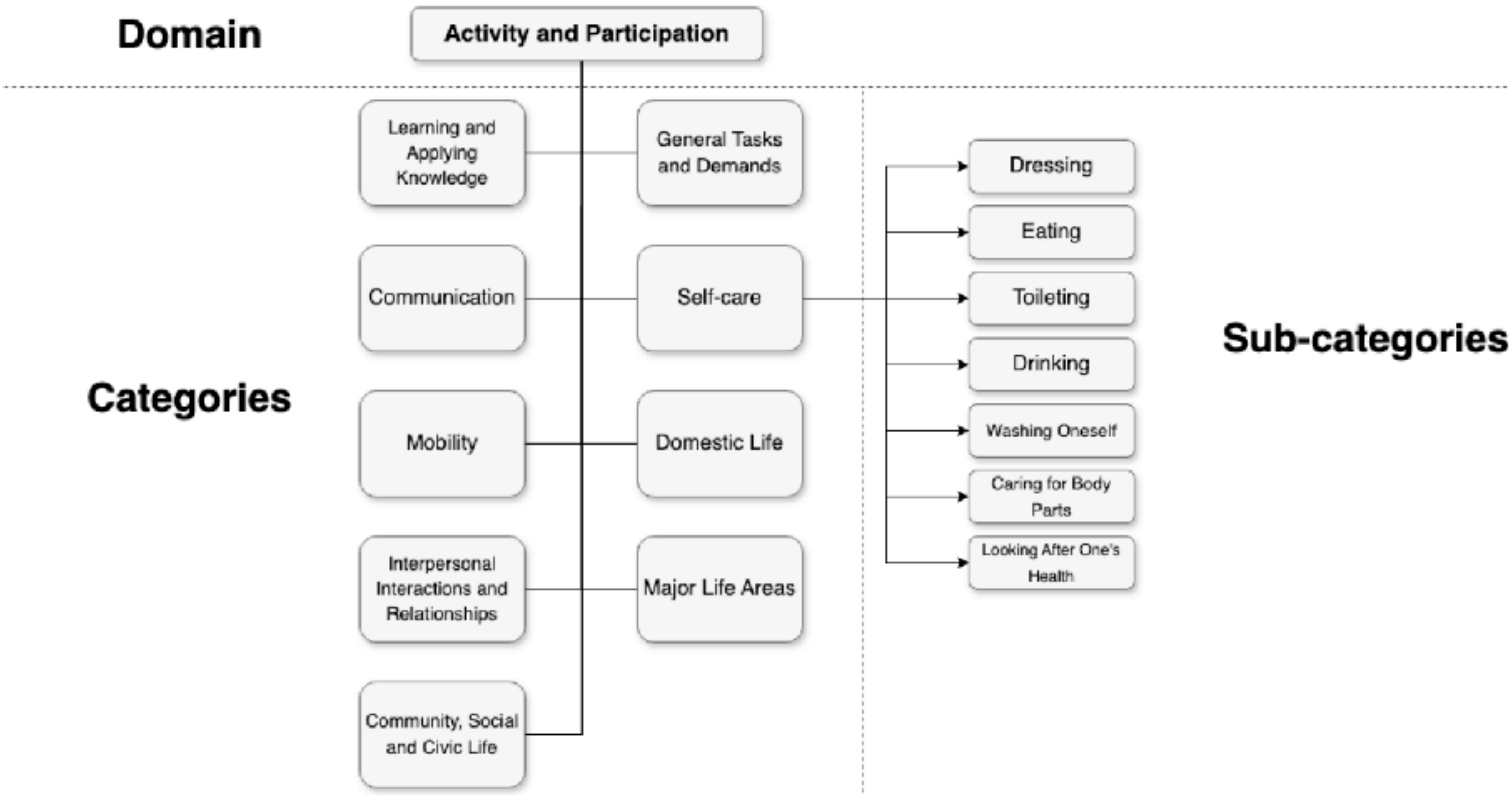


Figure 3.2: ICF Categories and Sub-categories

Generate conversations

```
query = [
    {"role": "system", "content": "Think as you are a real human."},
    {"role": "system", "content": """Generate one small and natural
    conversation without greetings by playing the roles of a friend (F)
    and an elderly patient (P). The conversation has 6 utterances,
    mentioning one or two events. Each utterance should be completed
    and has less than 20 tokens. The format is as below:
    F: utterance
    P: utternace
    F: utterance
    ...
    """},

    {"role": "user", "content": f"The topic of the conversation is
    about {category} events. In terms of functioning, {category} is
    about {definition}. {category} events include {events}."},

    {"role": "user", "content": f"The patient can talk with the
    friend about his/her daily life which can reflect his/her
    functioning in {category}. The friend should ask when and
    how the event occurs. "}
    ]
```

Figure 3.3: Prompt template for conversation generation

(1) Mobility
Chatbot: How have you been feeling lately? Have you been managing daily tasks well?
Patient: Well, I've had some trouble getting out of my chair to get into bed at night.
Chatbot: I see. When does that usually happen?
Patient: It usually happens when my back is acting up and I feel stiff.
Chatbot: Is there anything that makes it easier for you to transition?
Patient: Yes, using a grab bar next to the bed helps me to pull myself up.
Chatbot: That's great to know! How about other daily activities, are there any other occasions where mobility is challenging?
Patient: Yes, standing for long periods makes my legs feel weak and I need to sit down often.

(2) Domestic Life
Chatbot: How was your day today?
Patient: I spent the morning cleaning the house.
Chatbot: Do you clean the house every day?
Patient: No, I have a cleaning schedule for each day of the week.
Chatbot: That's quite organized. What do you do on other days?
Patient: On Tuesdays, I focus on doing the laundry and tending to the plants.
Chatbot: Sounds like a productive routine. How do you manage everything?
Patient: I've found that having a schedule helps me stay on track and manage my tasks efficiently.
Chatbot: That's great to hear. It must make things a lot easier for you.

(3) Self-care
Chatbot: How often do you usually do your skincare routine?
Patient: Oh, I do it every morning and night before bed.
Chatbot: That's great! How about brushing your teeth, when do you do that?
Patient: I brush my teeth after every meal and before bed.
Chatbot: Good habit. Do you find cutting your nails challenging?
Patient: Yes, it's a bit harder for me now. I do it every other week.
Chatbot: I see. What about caring for your general health? How do you manage that?
Patient: I take my medication daily and go for regular check-ups at the clinic.
Chatbot: That's important. Taking care of yourself is key.

Generate annotations

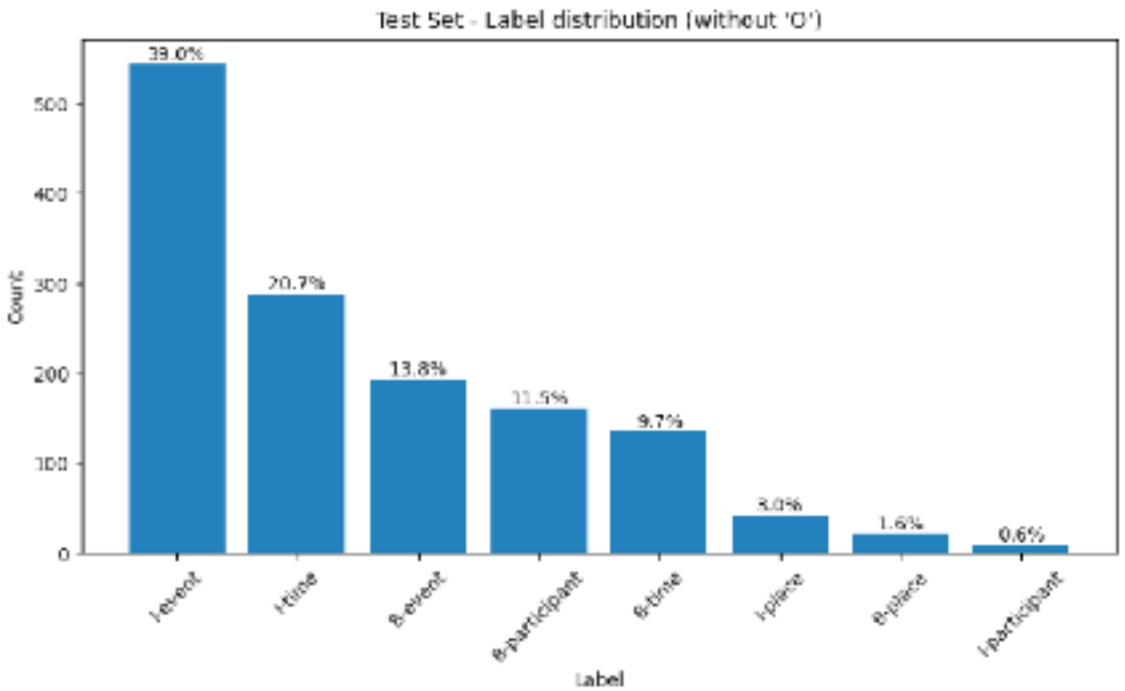
```
query = [
    {"role": "system", "content": "You are an expert in healthcare. I will provide you some conversations between\
     a chatbot and an elderly person. Please extract the information contained in each conversation: \
    'activity index', 'activity', 'participants', 'place', and 'time', and format every activity as a list."},
    {"role": "system", "content": "Each token in the conversation is provided with three information:\
     conversation id, sentence id, and token id."},
    {"role": "system", "content": "When extracting the information, please use the words and phrases appeared \
    in the original conversation. Please also indicate the sentence id and token id of the activity information."},
    {"role": "system", "content": f"The conversation index for this conversation is {conversation_index}. \
    If the conversation contains more than one activity, generate a list for each activity, using the\
     same conversation index but different activity indices. For a new activity, increment the activity index by 1."},
    {"role": "system", "content": "Generate the activity information only based on the conversation.\
     Do not use any external information."},
    {"role": "system", "content": "If there are no participants, place, or time of the activity mentioned\
     in the conversation, please mark as 'None' in the output."},
    {"role": "user", "content": f"Conversation: {con}"},
    {"role": "system", "content": "Please provide the output in the following JSON format:\
     [{'activity_index': 1, 'activity': 'activity 1', 'activity_sentence_id': 1, 'activity_token_ids': [1, 2], \
    'participants': 'participant 1', 'participants_sentence_id': 1, 'participants_token_ids': [3], \
    'place': 'place 1', 'place_sentence_id': 2, 'place_token_ids': [5], 'time': 'time 1', 'time_sentence_id': 3, \
    'time_token_ids': [7]}, {...}]. Please provide the output without Markdown code blocks, \
    and do not include the newline marker \\n in the output."},
]
```

Figure 3.7: Improved prompt template for generating activity information

Conversation data

Test data generated by GPT-Turbo 3.5 annotated by student:

478 utterances in 54 conversations



Training data generated by GPT-Turbo 3.5 and annotated by GPT-4o

Category	Raw Data Conversation	Raw Data Utterance	Cleaned Data Conversation	Cleaned Data Utterance
Mobility	280	2765	241	2049
Self-care	420	4312	335	2944
Domestic Life	240	2413	219	1921
Total	940	9490	795	6914

Table 3.1: Dataset by Category

Chuqiao Guo, Text Mining Master Thesis 2024

Performance results

Model	Precision	Recall	F1-score
Rule-based	0.30	0.49	0.30
Fine-tuned BERT	0.68	0.62	0.62
GPT-Prompting	0.68	0.73	0.70

Table 6.1: Model Performance Comparison

	Precision	Recall	F1-Score	Support
B-event	0.16	0.79	0.26	193
B-participant	0.31	0.94	0.46	161
B-place	0.05	0.68	0.09	22
B-time	0.16	0.30	0.21	135
I-event	0.28	0.30	0.29	544
I-participant	0.30	0.33	0.32	9
I-place	0.15	0.36	0.21	42
I-time	0.45	0.06	0.10	288
O	0.87	0.63	0.73	4801
Accuracy			0.58	6195
Macro Avg	0.30	0.49	0.30	6195
Weighted Avg	0.74	0.58	0.62	6195

Table 6.2: Classification Report - Rule-based System

	Precision	Recall	F1-Score	Support
B-event	0.67	0.67	0.67	193
B-participant	0.78	0.09	0.16	161
B-place	0.22	0.18	0.20	22
B-time	0.84	0.79	0.82	135
I-event	0.63	0.68	0.66	544
I-participant	0.64	0.78	0.70	9
I-place	0.56	0.52	0.54	42
I-time	0.86	0.90	0.88	288
O	0.92	0.94	0.93	4801
Accuracy			0.88	6195
Macro Avg	0.68	0.62	0.62	6195
Weighted Avg	0.87	0.88	0.87	6195

Table 6.3: Classification Report - Fine-tuned BERT System

	Precision	Recall	F1-Score	Support
B-event	0.68	0.69	0.68	193
B-participant	0.72	0.75	0.73	161
B-place	0.44	0.64	0.52	22
B-time	0.80	0.76	0.78	135
I-event	0.70	0.78	0.73	544
I-participant	0.57	0.44	0.50	9
I-place	0.46	0.74	0.56	42
I-time	0.81	0.82	0.81	288
O	0.95	0.93	0.94	4801
Accuracy			0.89	6195
Macro Avg	0.68	0.73	0.70	6195
Weighted Avg	0.90	0.89	0.89	6195

Table 6.4: Classification Report - GPT-Prompting System

Chuqiao Guo, Text Mining Master Thesis 2024

Thank you!