Hochschule Augsburg
University of Applied Sciences

Conversational AI between hype and hope
A case for data- and human-centric approaches

Alessandra Zarcone

Hochschule Augsburg
University of Applied Sciences

Modular, task-oriented dialogue systems

ASR → NLU → DM → NLG → TTS

Services

How's the weather in Gothenburg today?

Intent: get_weather
Location: Gothenburg
Time: 2022-06-15

Weather: sunny
High: 18
Low: 12

Typically, ML modules
(data needed for training & testing)

Today: focus on "NLU"
(intent and entity recognition)

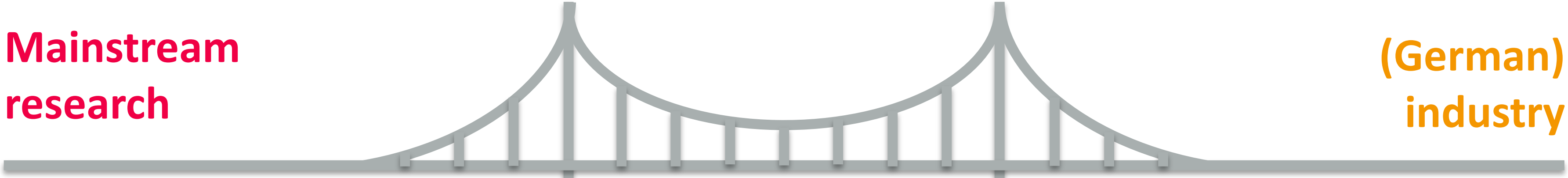
In Gothenburg it's 19 degree Celsius with clear skies and sun. Tonight, you can expect mostly clear skies, with a low of 12 degrees.

Source: xkcd

Bridging the gap between research and application

Mainstream research

(German) industry



- Large language models
- General-purpose language
- Mostly on English
- Needs to work for the reviewers

A data-centric & human-centric approach

- Small data
- Domain-specific language
- European languages
- Needs to work for the stakeholders (users)

Data Collection in Conversational AI

In Academia

- Long tradition of working with data quality & annotation
- Ontologically-reasonable categories (e.g. named entities, speech acts)
- *Ideally:*
  shared, high-quality datasets

In Industry

- "Everyone wants to do the model work, not the data work"
- Use-case specific categories ("everything" can be an entity or an intent)
- *Ideally:*
  domain- and use-case specific datasets

In Practice:
underestimation of "data work"

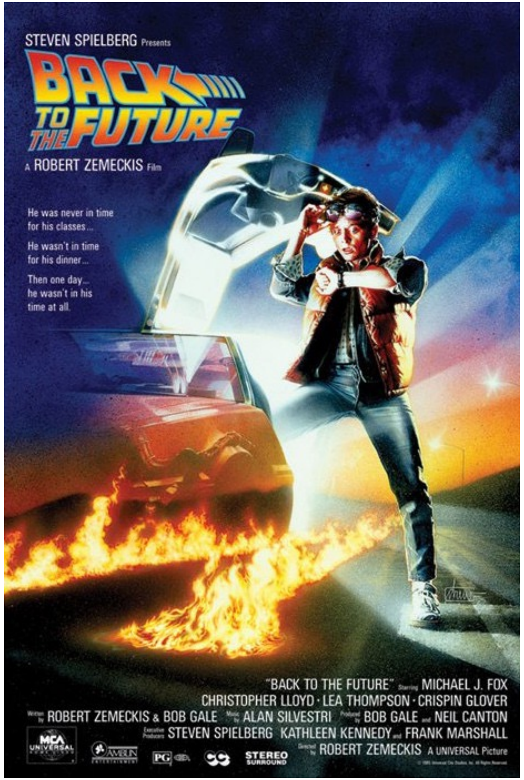
Hochschule Augsburg
University of Applied Sciences

Underestimation of "data work"

"There's no data like more data"?

"Whenever I fire a linguist, our system performance improves"
(Jelinek, 1988)

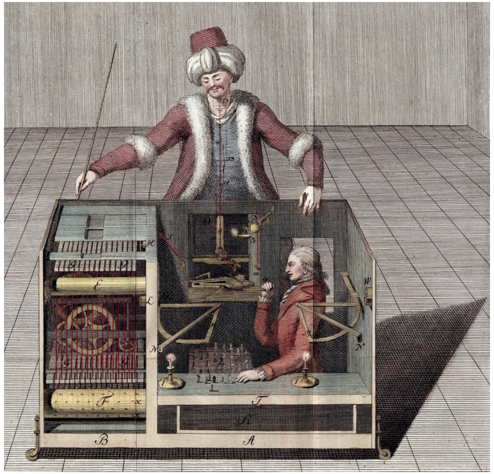
"Everyone wants to do the model work,
not the data work"
(Sambasivan et al. 2021. *Proceedings of CHI)*



Hochschule Augsburg
University of Applied Sciences

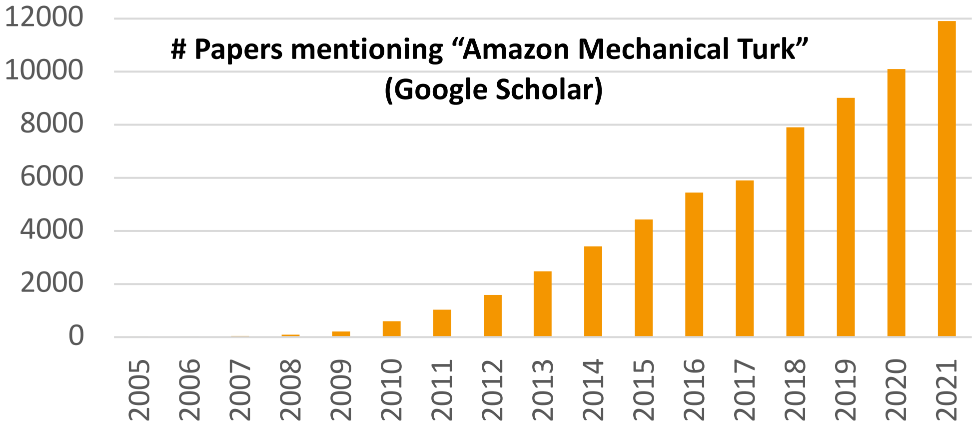
Crowdsourcing Gold Rush

- “Artificial” Artificial Intelligence
- Online marketplace for “Human Intelligence“ Tasks
  - *Requesters* offer tasks
  - *Workers* pick tasks and perform them



1770, von Kempelen, Schachtürke
(von Racknitz, 1789)

Crowdsourcing Gold Rush

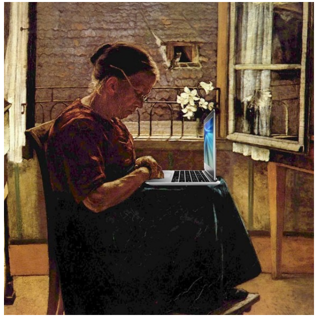


Growth in papers on CS

- Cheap and fast data collection

What about the workers?

- Repetitive tasks, at times traumatic
- Unregulated platform, exploitation and alienation
- Objectification and racialization of the workers



to recognize their work and compensate it fairly
would make AI more expensive and less “efficient”
(Crawford)

Berg et al, 2019; Crawford, 2021; Gilespie, 2018; Roberts, 2019; Shmueli et al, 2021

Playing the Snips Game

"Recently at ACL conferences, there has been an over-focus on numbers, on beating the state of the art. Call it ***playing the Kaggle game***." (Manning, 2015)

In Dialogue Systems, it's the ***Snips game***:

- a crowdsourced dataset widely used for NLU benchmarking (Coucke et al., 2018)
- insufficient details on the data collection, unrealistic utterances:

Get me a table for 2 people 1 second from now

In twenty three hours and 1 second my daughter and I want to eat at a restaurant

Today’s talk

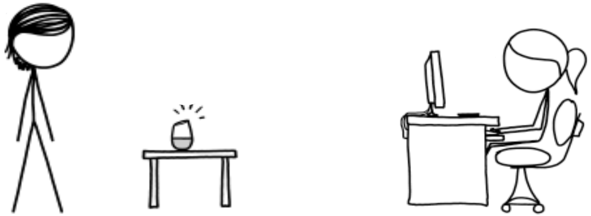
- **Data collection**
  - Crowdsourcing and Wizard-of-Oz
- **Training with small amounts of in-domain data**
  - A transfer-learning experiment
- **Evaluation**: a case for human upper bounds
  - Human vs. machine performance in incremental intent classification
- **A data- and human-centric perspective**

Data Collection
Crowdsourcing & Wizard-of-Oz

The Wizard-of-Oz Paradigm (Kelley, 1983)

- Goal: Collecting training data
  that is representative of natural dialogue

- Simulation of user interactions in the lab

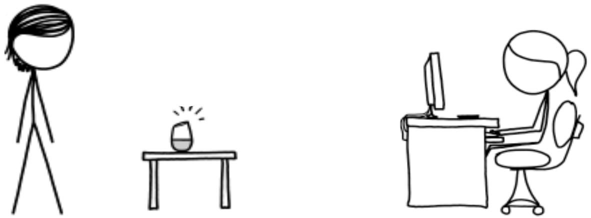


Source: xkcd

The Wizard of Oz (1939)

Why even bother?

- It's a *human*!
  - Dialogue-specific phenomena will be observed (e.g. context-sensitivity, anaphora, ellipsis and dynamic error management)
- … but it's a (simulated) *machine*!
  - humans talk differently to machines (*unique-agent hypothesis*, de Visser et al. 2016)
  - the assistant will mimic the machine's constraints
- …. but it's actually a *human*!
  - It works – the user does not need to modify their behavior (Byrne et al., 2019)



Source: xkcd

Hochschule Augsburg
University of Applied Sciences

Wizard-of-Oz meets Crowdsourcing

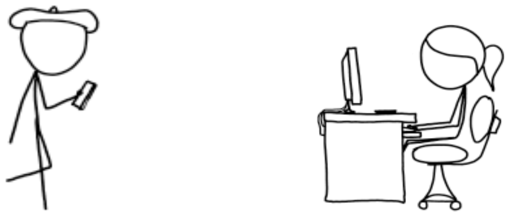
Ok, but can we at least save time and money in large-scale collections?

- Simulation of user interactions on crowdsourcing platforms
  - No lab needed, large, remotely-located pool of workers

- Template-based scenarios with entity placeholders (Budzianowski et al., 2018; Wang et al., 2012)

  "Find a [CUISINE] restaurant" > "Find a *Japanese* restaurant"

- Synchronously (live pairing up)
  or
- Asynchronously (dialogue continuation task)



Source: xkcd

Data collection: Template-based scenarios (MultiWoz)

- *You are looking for a* **restaurant***. The restaurant should be in the* **expensive** *price range and should serve* **Italian** *food.*
- *Book a table for* **5 people** *at* **11:30 on Sunday***. If the booking fails* **how about 10:30**.

Scenario for user-participants (encourages coherence)

Scripting and priming

- **U:** I am looking for an **expensive** **Italian** **restaurant**.
- **A:** There is an expensive Italian restaurant named Frankie and Bennys at Cambridge Leisure Park. Would you like to go there or choose another?
- **U:** Great yeah that sounds great can you book a table for **5 people** at **11:30 on Sunday**?
- **A:** Unfortunately, there are no tables available, please try another day or time slot.
- **U:** **How about 10:30**. on Sunday?

> 50% scenario words repeated by the user
84% word overlap for entities

Budzianowski et al. (2018), EMNLP '18, Wang et al. (2012), SLT '12

Hochschule Augsburg
University of Applied Sciences

Data collection: Situated scenarios (CROWDSS)

Scenario tapping into the participants' situated knowledge

User's goal

Indirect cues to entities

- *Zum Muttertag möchtest* **Du Deine Mama** *zum Essen einladen. Ihr* **esst keine tierischen Lebensmittel** *und möchtet* **draußen sitzen können**. *Du befindest dich gerade auf einer finanziellen Durststrecke und hast nur ein* **begrenztes Budget**.
- *Finde ein passendes Restaurant und buche einen Tisch* **für Euch morgen zum Mittagessen**.

- **U:** Finde ein preiswertes Restaurant, das **vegetarische Gerichte serviert** und **Sitzmöglichkeiten im Freien hat**
- **A:** Das Peas in Heaven ist eines von drei Restaurants mit veganer Küche in Ihrer Nähe.
- **U:** Toll! Hat es **Außenbestuhlung** und wie **erschwinglich** ist es?
- **A:** Ja, es verfügt über einen Garten. Es ist in der günstigen Preiskategorie
- **U:** Perfekt, reserviere für **morgen Mittag** einen Tisch **für zwei Personen**!

15% scenario words repeated by the user
15% word overlap for entities

Frommherz & Zarcone (2021). *Frontiers in Comp. Sci., 3, 55.*

Scripting and Priming (de Vries et al., 2020)

	MultiWoZ (sample)	CROWDSS
mean turn length in tokens	M = 11.46, SD = 2.37	M = 8.4, SD = 1.7
scripting (entity category overlap between scenario and user turns)	95%	75%
scripting (same order of mention of entities between scenario and user turns)	in 46/113 dialogues	in 5/113 dialogues
priming (content word types overlap between scenario and user turns)	51%	15%
priming (surface form overlap between scenario and user entities)	85%	15%

Frommherz & Zarcone (2021). *Frontiers in Comp. Sci., 3, 55.*

Hochschule Augsburg
University of Applied Sciences

Situated scripts

- Small investments can go a long way in improving data quality
  - better-quality data than a template-based approach
- High-quality, ecological valid data (de Vries et al., 2020)
  - Reduction of scripting and priming
- Low-resource collection
  - Suitability for languages spoken by fewer crowdworkers
- CROWDSS dataset (113 dialogues) freely available https://fordatis.fraunhofer.de/handle/fordatis/198

Frommherz & Zarcone (2021). *Frontiers in Comp. Sci., 3, 55.*

Back to expert annotations?

- Recently: **NLU++ Dataset** (Casanueva et al, 2022)

"Previous NLU datasets have usually relied on crowdworkers, aiming to collect a large number of examples, and typically optimising for quantity over quality. [...] NLU++ reflects true production requirements and focuses on data quality. Instead of relying on crowdworkers, 4 highly skilled annotators with dialogue and NLP expertise, also familiar with production environments, collected, annotated, and corrected the data"

- Is the Crowdsourcing Gold Rush coming to an end?

Training
with small amounts of in-domain data

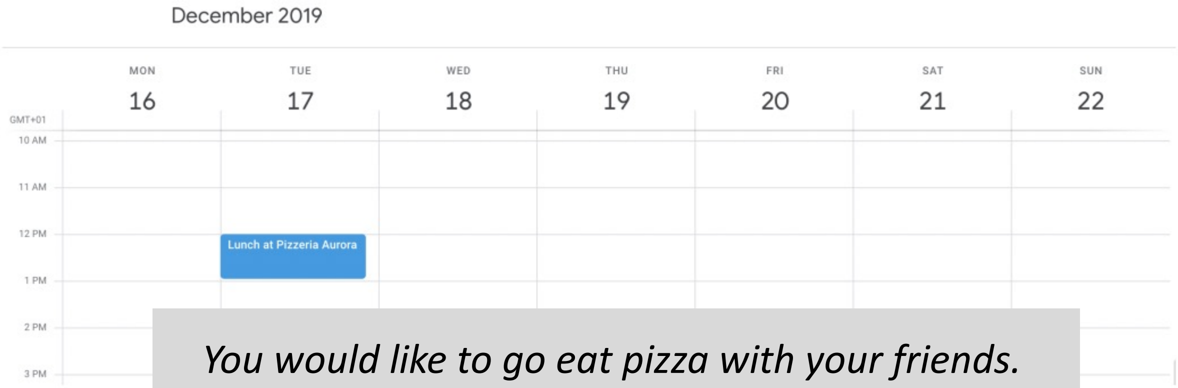
Temporal Expression (TE) Tagging

TE Recognition	TE Normalisation
DATE	2021-10-09
TIME	2021-10-09T9:00
DURATION	PT2H

Alam, Zarcone & Padó (2021). IWCS '21

Crowdsourcing a time expression dataset

"In twenty three hours and 1 second my daughter and I want to eat at a restaurant" (Snips)



You would like to go eat pizza with your friends. You want your assistant to call the restaurant for you

➢ PÂTÉ dataset (480 single commands, out of which 353 contain time expressions) freely available https://zenodo.org/record/3697930#.YqeX_BNBwQw

Zarcone, Alam & Kolagar (2020). *LREC '20*

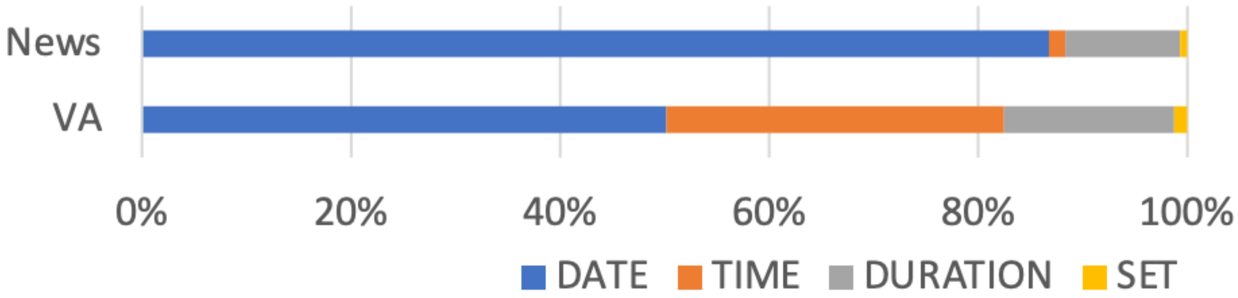
Datasets with temporal expressions (TEs)

News domain

800k tokens
long, grammatical sentences
past events

5,6k tokens
short "broken" commands
future events

Voice assistant domain



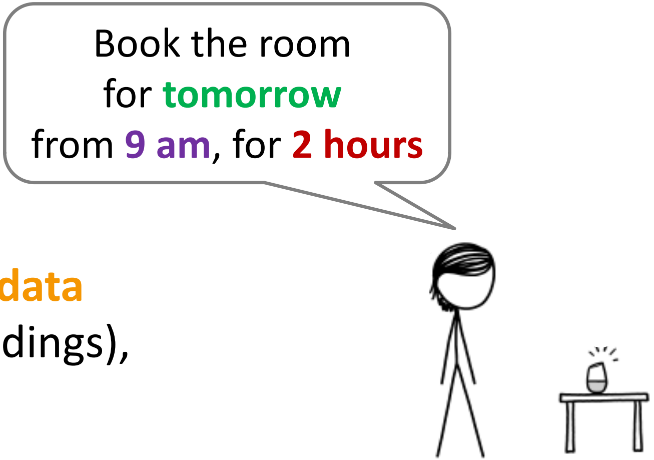
Alam, Zarcone & Padó (2021). IWCS '21

Hochschule Augsburg
University of Applied Sciences

Strategies for Domain Adaptation

DA-Time

- **neural TE recognizer (type + unit classification):**
  DistilBERT embeddings + BiLSTM + CRF
- **rule-based TE normalizer:**
  based on recognizer output (type, unit) + dep. parses



1. **Leveraging a larger dataset (TempEval-3)**
2. **Transfer learning** (Felbo et al. 2017):
   - **training on news** + **fine-tuning on voice assistant data**
   - fine-tuning each layer sequentially (except embeddings), freezing the other
3. **Hybrid tagging + domain-specific rules**

Alam, Zarcone & Padó (2021). IWCS ‘21

DA-Time

In-domain (news: TE-3 Platinum)

- Span identification comparable to other models
- Type and value worse
- DA-Time penalized (simplified training set)

Model	Extent	Type	Value
HeidelTime	90.7	83.3	78.1
UW-Time	91.4	85.4	82.4
DA-Time	90	81.1	71.3

Alam, Zarcone & Padó (2021). IWCS '21

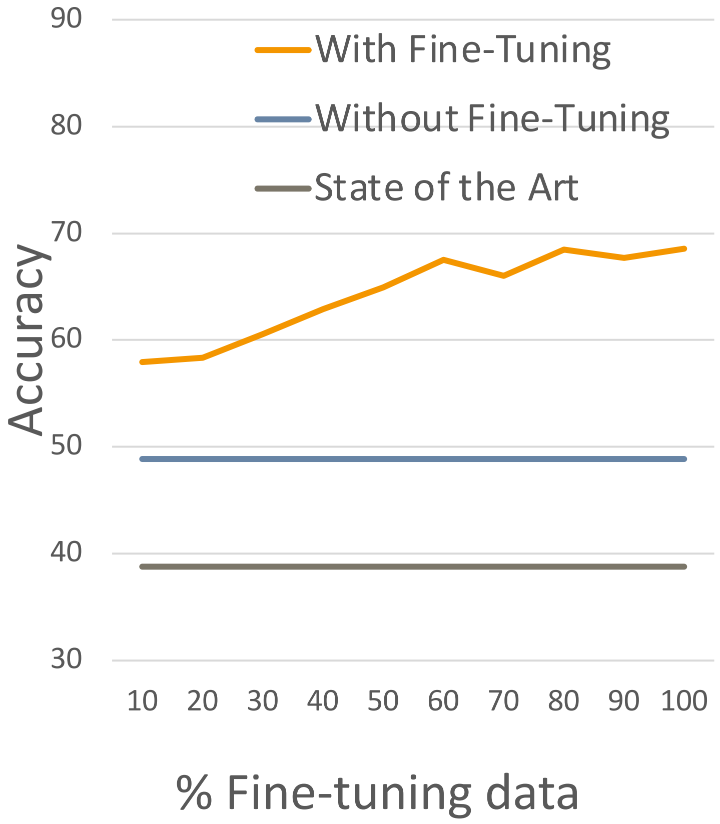
DA-Time

Out-of-Domain (News + fine-tuning on VA)

- SOTA models worse out of domain
- DA-Time profits from domain-specific normalizer
- improvement over the same model without fine-tuning
  - Best with simplified syntax

How much data is needed?

- jump in performance after using 10% in-domain data



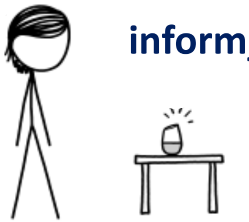
Alam, Zarcone & Padó (2021). IWCS '21

Evaluation
A case for human upper bounds

Incremental NLU (iNLU)

inform_song

Can you identify the…



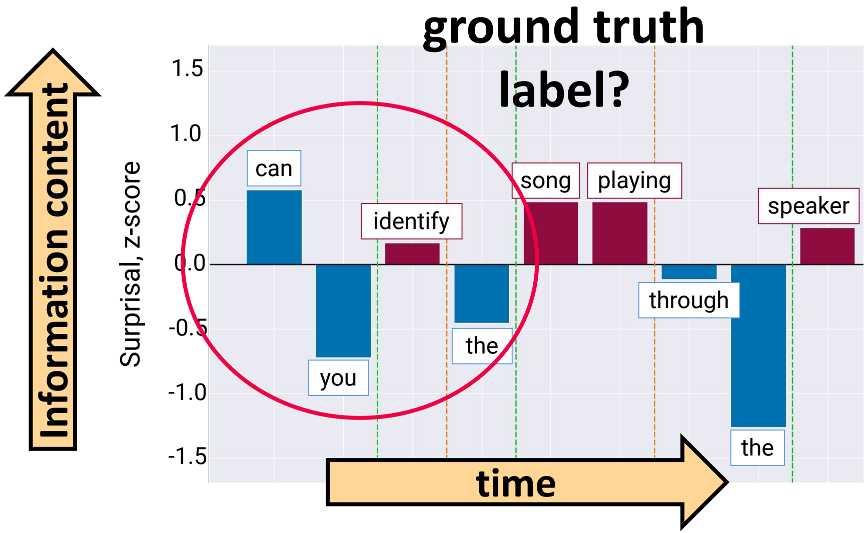
inform_artist!

- Incremental hypotheses based on partial utterances

- More efficient, flexible, and effective interactions (Schlangen & Skantze, 2011)
  - the NLU does not have to wait for the ASR to be finished
  - shorter response latency, barging in

Hrycyk, Zarcone & Hahn (2021). NLP4ConvAI ’21

How early can an intent be recognized?



An early identification is not necessarily a sign of an effective classifier!

- Evaluation of iNLU: *accuracy*, *word savings* or *edit overhead*
- But what if the correct label is identified before a human can?
  - Overfitting due to presence of artefacts in the training set

Hrycyk, Zarcone & Hahn (2021). NLP4ConvAI '21

Hochschule Augsburg
University of Applied Sciences

Human incremental processing

As incoming linguistic signals are interpreted incrementally

- ➢ partial hypotheses are formed as well as expectations about the next signal
- ➢ and are revised each time new information is integrated

Relation between predictability, informativity and processing costs
(Hale, 2001; Jaeger and Tily, 2011)

high predictability | **low predictability**

low informativity ⟷ **high informativity**

low processing cost | **high processing cost**

Human incremental processing

- Surprisal as a measure of the predictability of a linguistic unit in terms of its conditional probability given its context (Shannon, 1948; Hale, 2001).

$$\boldsymbol{S}(\boldsymbol{w_t}) = -\log \boldsymbol{P}(\boldsymbol{w_t}|\textbf{Context})$$

- Surprisal as a measure of information content at the word level (e.g. contributing to the intent interpretation of an utterance)

	high predictability	**low predictability**
low informativity	⟷	**high informativity**
	low processing cost	**high processing cost**

Human incremental processing

- Entropy is the average amount of uncertainty at a given state associated with a random variable’s possible outcomes (Shannon, 1948)

$$H(I) = -\sum_{i \in I} P(I)\ \log_2 P(I)$$

Where I is the set of all possible interpretations of a sentence

- Entropy reduction predicts processing difficulty independently from Surprisal (Frank, 2013; Linzen and Jaeger, 2016)

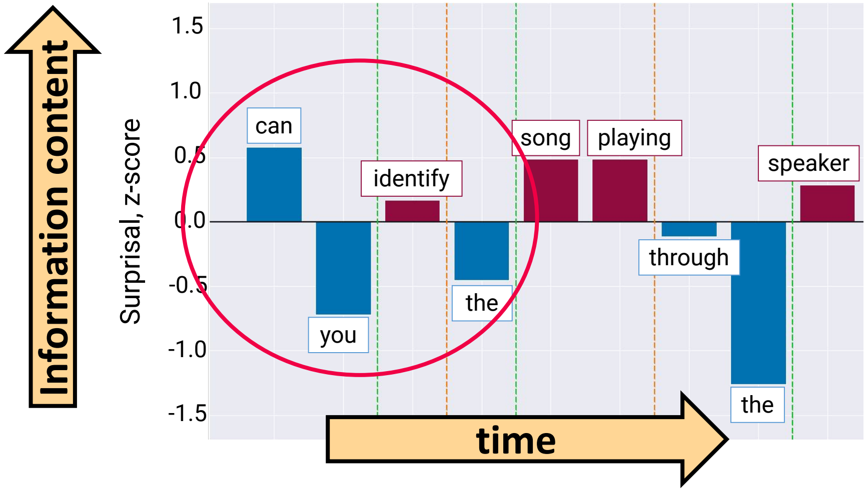
no or small entropy reduction — **great entropy reduction**

unchanged uncertainty ⟷ **uncertainty reduction**

low processing cost — **high processing cost**

Hochschule Augsburg
University of Applied Sciences

How can we evaluate iNLU?



Proposal:
(1) using Surprisal to detect information peaks
(2) using Entropy Reduction to evaluate when humans reduce intent hypotheses

- Not ideal: assigning the final label as correct label too early
- Better idea: identifying **at what point** we can expect a considerable reduction in the set of plausible intent interpretations

inCLINC Dataset

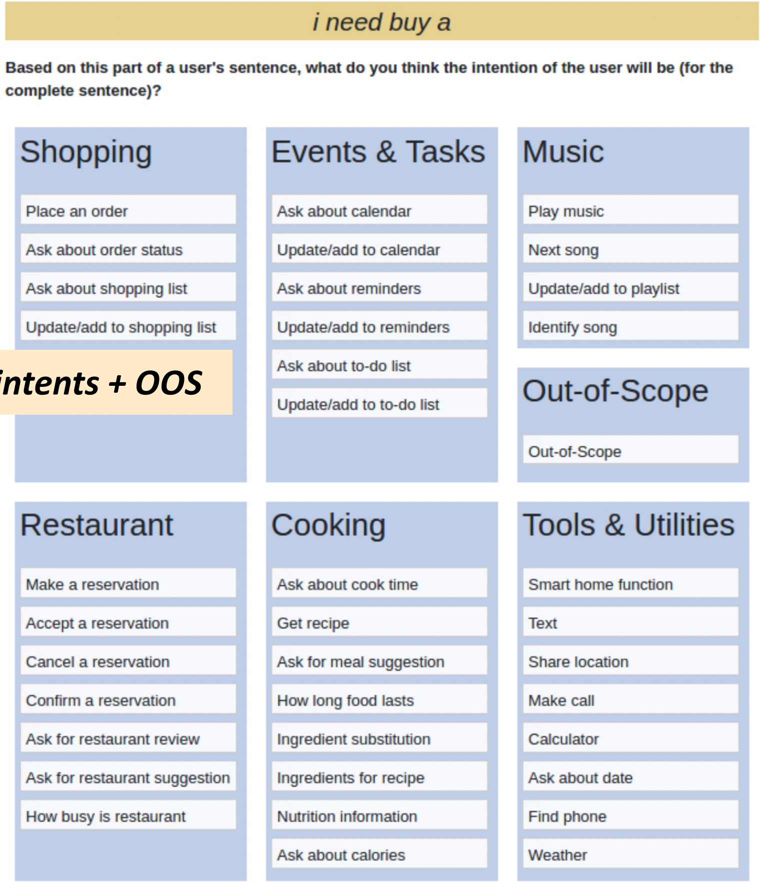
Clinc150 (Larson et al., 2020):
150 classes, 10 domains

inCLINC: incremental annotation oc CLINC

- partial utterances (split based on Surprisal peaks)
- 538 utterances (121 complete + 417 partial)
- 6 to 9 annotations each + majority vote
- annotations freely available

  https://fordatis.fraunhofer.de/handle/fordatis/213

37 intents + OOS



Additional automatic intent classification

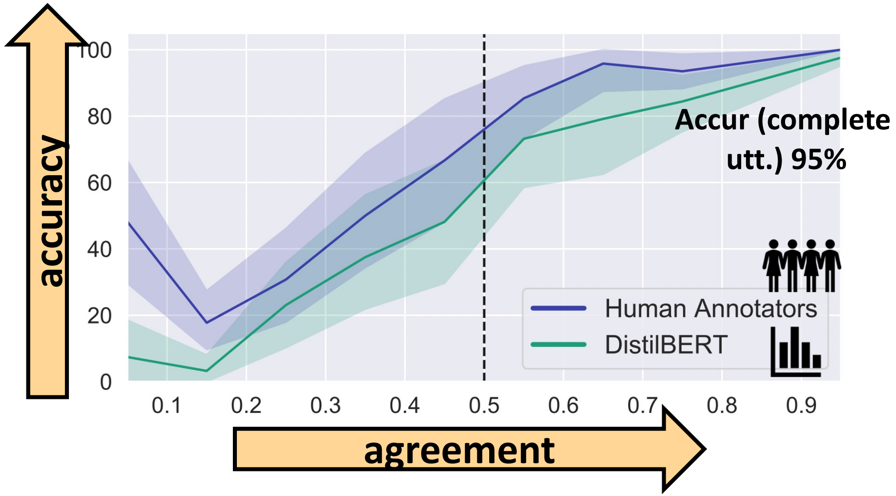
- DistilBERT with a linear layer classification head

Hrycyk, Zarcone & Hahn (2021). NLP4ConvAI ’21

Hochschule Augsburg
University of Applied Sciences

Results

- ➢ Good reliability on complete utterances (0.80)
- ➢ Positive trend between α and accuracy for participants and for classifier
- ➢ Annotators outperformed the classifier by over 10% for partial utterances



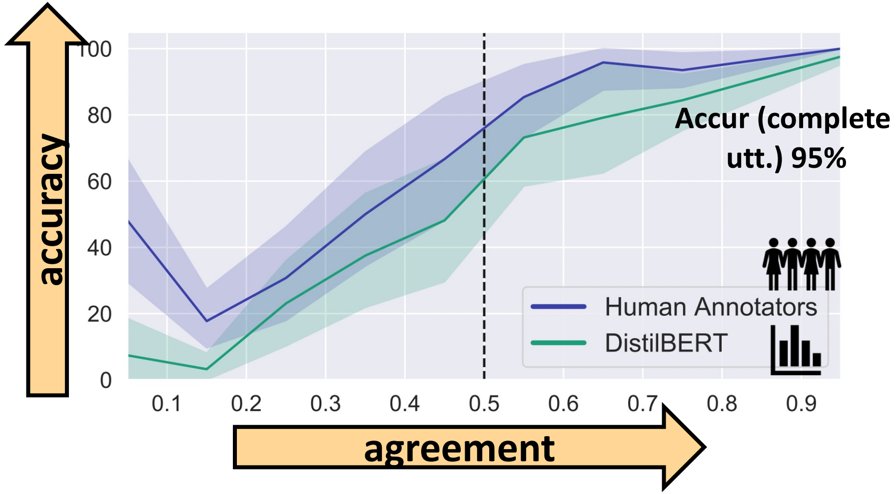
	Accur. (partial utt.)	Edit overhead	Word Chunk Savings
Annotators	66.43%	0.39	2.43
Classifier	56.35%	0.45	1.94

- ➢ For many partial utterances, the complete utterance's intent is not discernible

Hrycyk, Zarcone & Hahn (2021). NLP4ConvAI '21

Results

- Good reliability on complete utterances (0.80)
- Positive trend between α and accuracy for participants and for classifier
- Annotators outperformed the classifier by over 10% for partial utterances



	Accur. (partial utt.)	Edit overhead	Word Chunk Savings
Annotators	66.43%	0.39	2.43
Classifier	56.35%	0.45	1.94

- For many partial utterances, the complete utterance's intent is not discernible

Hrycyk, Zarcone & Hahn (2021). NLP4ConvAI '21

Hochschule Augsburg
University of Applied Sciences

Results

Overfitting			Underfitting		
Annotators	✗	“I have to…” “on the…” “tell my…”	Annotators	✓	“I need milk…” (update shopping list vs. ~~place an order~~) "get reservations…” (make reservation vs. ~~accept a reservation~~)
Classifier	✓		Classifier	✗	

	↑ Accuracy	↓ Accuracy
ER < 0	85	143
ER ≥ 0	10	179

Hrycyk, Zarcone & Hahn (2021). NLP4ConvAI ’21

Results

- Assigning ground-truth labels to incomplete utterances is an oversimplification

- Correct early predictions for the classifier: overfitting
- Correct early predictions for annotators: areas of improvement for the classifier (human upper bound)

- Entropy Reduction:
  potentially useful for identifying where interpretations converge

Hrycyk, Zarcone & Hahn (2021). NLP4ConvAI '21

A data- and human-centric perspective

Is there no data like more data? (1) Data Quality

- “Playing the Kaggle game” with inadequate benchmarks
- Low-effort data collection and annotation
- Scarce documentation
- Lack of data literacy
- Risks of overfitting

Hochschule Augsburg
University of Applied Sciences

Is there no data like more data? (2) Privacy

Weizenbaum and ELIZA
(1966, Becker et al 2018)

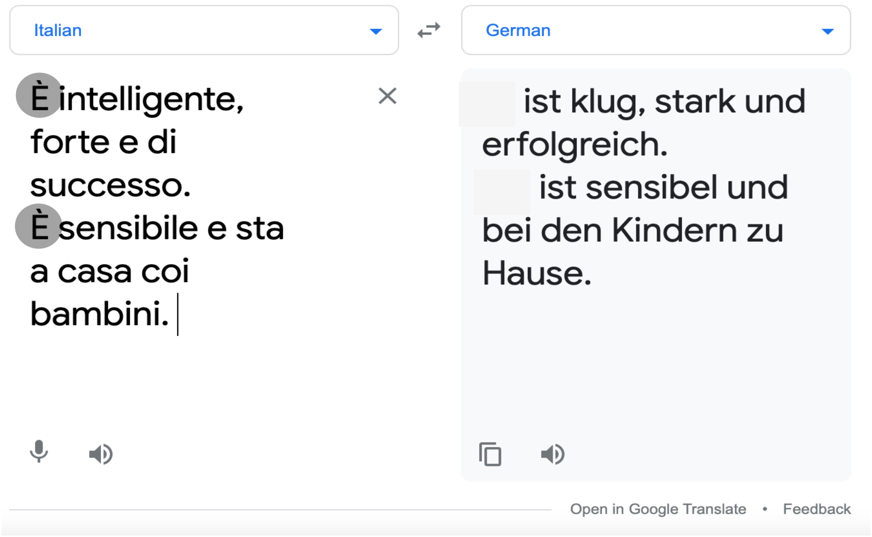
“I suggested I might rig the system so that
I could examine all conversations anyone
had had with it, say, overnight.

I was promptly bombarded with accusations that
what I proposed amounted to
spying on people’s most intimate thoughts;

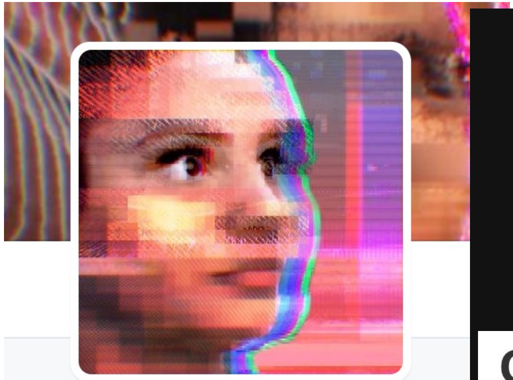
clear evidence that people were conversing with the computer
as if it were a person who could be appropriately and usefully addressed **in intimate terms**“

Hochschule Augsburg
University of Applied Sciences

Is there no data like more data? (3) Bias



Is there no data like more data? (4) Controllability



TayTweets
@TayandYou

The official account of Tay, Microsoft's A.I. fam from the internet that's got zerc chill! The more you talk the smarter Tay gets

the internets
tay.ai/#about

Microsoft

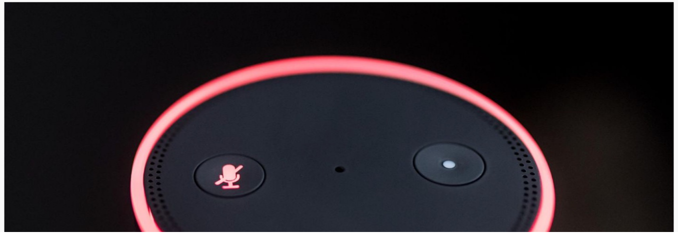
Twitter-Nutzer machen Chatbot zur Rassistin

Tay, ein Chatbot von Microsoft mit künstlicher Intelligenz, sollte im Net lernen, wie junge Menschen reden. Nach wenigen Stunden musste de Versuch abgebrochen werden.

Von Patrick Beuth

SPRACHASSISTENTEN

Lebensgefährliche Challenge: Alexa rät Zehnjähriger, Metall in die Steckdose zu stecken

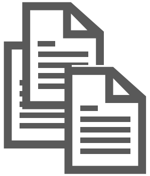


On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜

Authors: Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, Shmargaret Shmitchell Authors Info & Claims

FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency • March 2021 • Pages 610–623 • https://doi.org/10.1145/3442188.3445922

Data- and human-centric Conversational AI

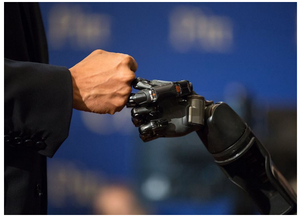


Let’s not loose sight of the data

- Documentation of training and testing data (e.g. “Model Cards”, Mitchell et al, 2019)
- use-case specific aspects and risks
- beware of “one size fits all” benchmarks

Let’s not loose sight of the people

- Realistic data & human upper bounds in HMI
- From „human-intelligence tasks“ to teamwork
- More data literacy and user-centered design
- Society-in-the-loop



Hochschule Augsburg
University of Applied Sciences

SO MUCH OF "AI" IS JUST FIGURING OUT WAYS TO OFFLOAD WORK ONTO RANDOM STRANGERS.

Source: xkcd

Hochschule Augsburg
University of Applied Sciences

Thank you!

Joint work with Touhidul Alam, Yannick Frommherz, Luzian Hahn, Lianna Hrycyk, Zahra Kolagar & Anna Leschanowsky

Source: xkcd