Dialogue Modelling

Alberto Testoni & Esam Ghaleb
Institute for Logic, Language and Computation (ILLC)
Dialogue Modelling Group (DMG)

NLP1 guest lecture, December 2023
(Credits to prof. Raquel Fernández for some of the slides)

Plan for today

Part 1 (Alberto Testoni):

- What is Dialogue Modelling?
- Current NLP Methods to Model Dialogue Systems / Chatbots
- The challenges of Multimodal Visual & Language Dialogue systems

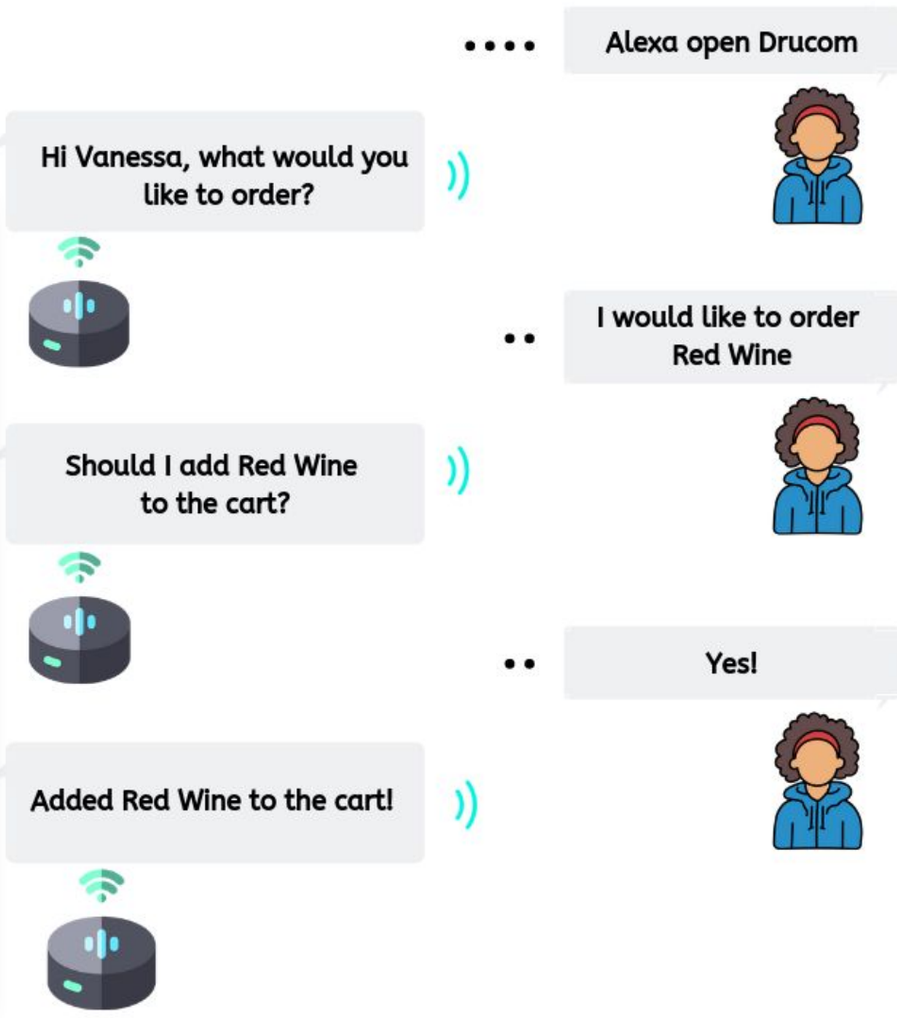
Part 2 (Esam Ghaleb):

- Multimodality in Face-to-Face Dialogues (Gestures & Speech)

Dialogue

What is it and why do we care

- Using language for cross-speaker communication and interaction
- Primary form of language use and language learning



Dialogue

What is it and why do we care

It is convenient to distinguish between

- Social chit-chat dialogue

A: What's your favorite holiday?
B: I'm a big fan of Christmas.
A: Is that so? Mine is Holloween.
B: I also like Holloween. But I like Christmas most.

- Task-oriented dialogue

PC: Alexa, open plan my trip.
ALEXA: Where are you planning to go?
PC: I'm going to Portland.
ALEXA: What city are you leaving from?
PC: Seattle.
ALEXA: What date are you flying out?
PC: Next Thursday.
ALEXA: This will be fun. You go from Seattle to Portland on April 27th, 2017.

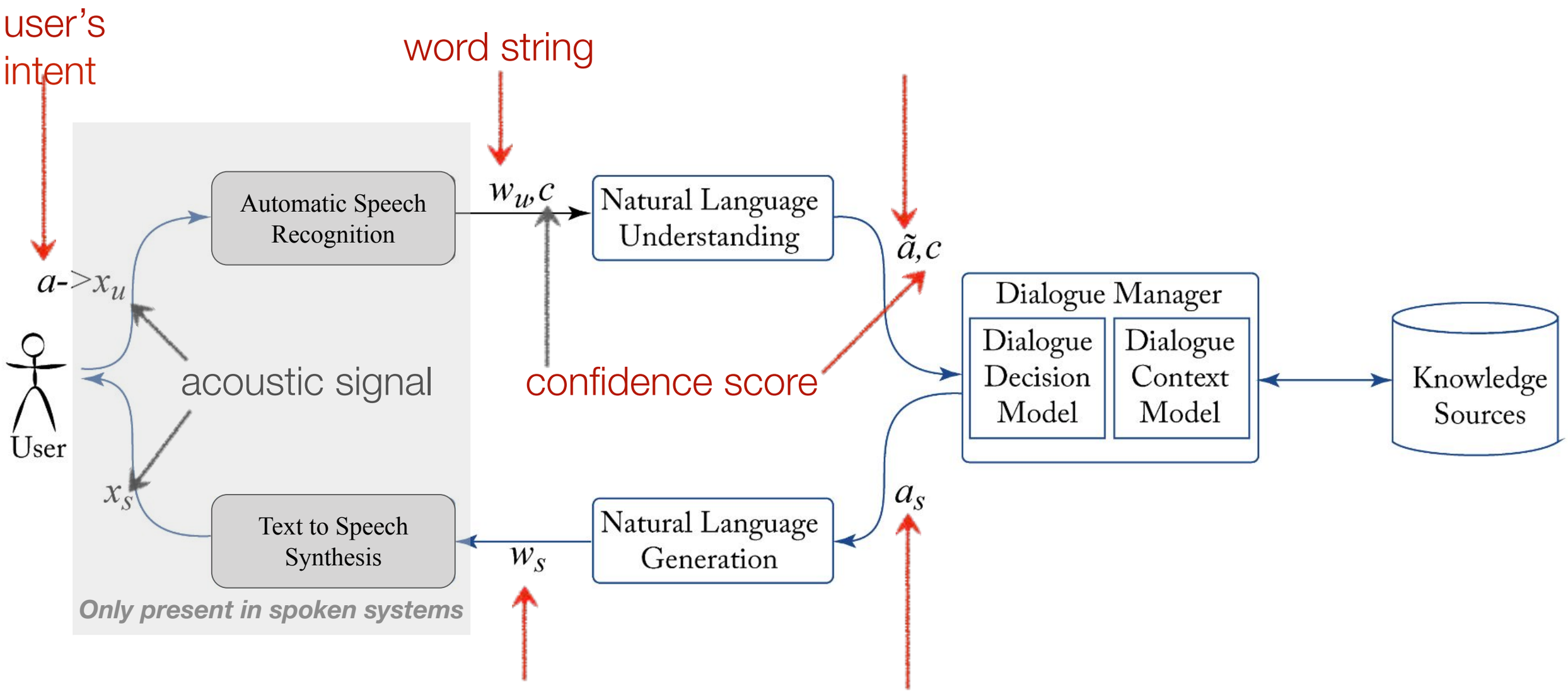
Dialogue modelling

Modelling a dialogue agent involves:

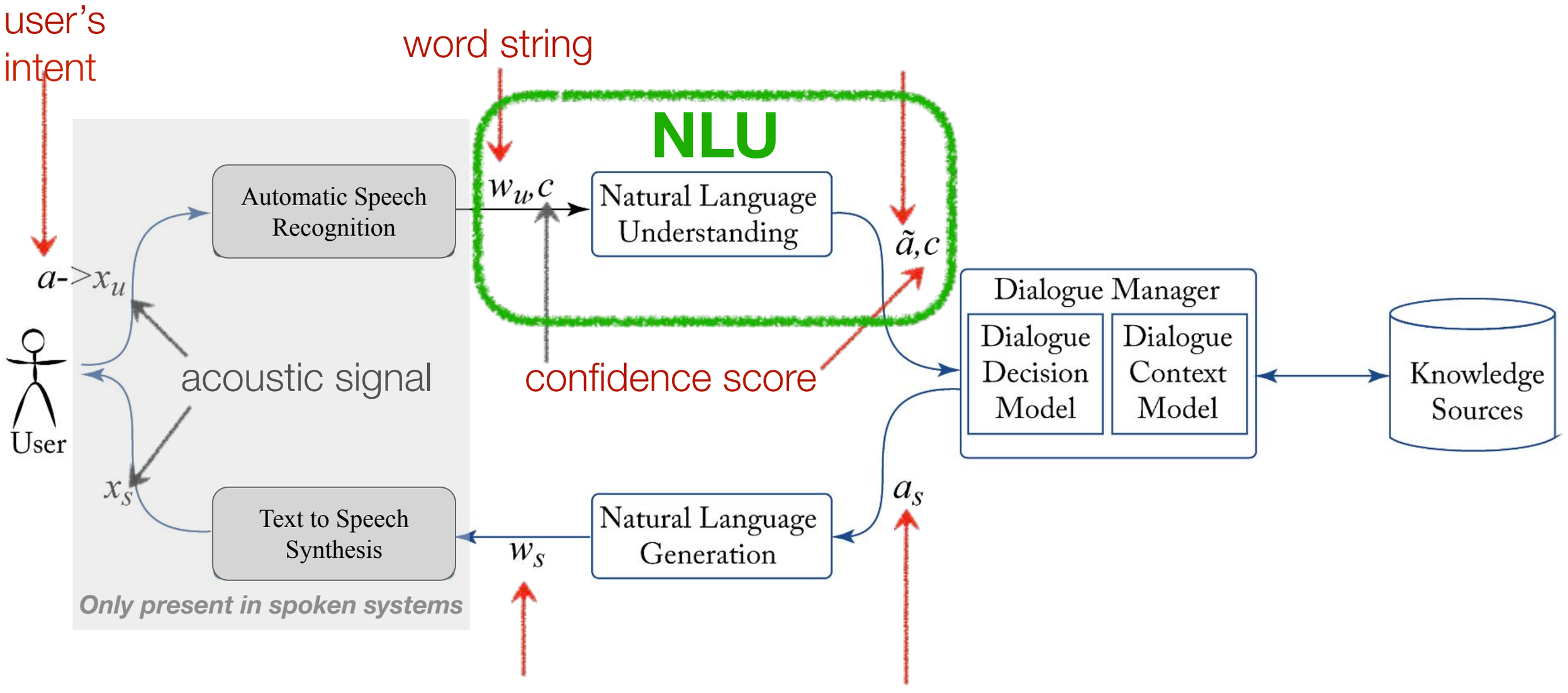
- Understanding the utterances by the dialogue partner.
- Keeping track of the dialogue history.
- Deciding what to say.
- Generating an utterance that conveys the speaker’s intend.

A dialogue agent (McTear, 2020)

- Task-oriented dialogue agents are typically modelled using a **modular architecture**, with modules for the steps above



A dialogue agent (McTear, 2020)



NLU

Intent prediction: Why is it difficult?

Speech act or **dialogue act**: the function of (or the action performed by) an utterance. The intention of the speaker.

- *statement, question, answer, agreement, request*, ….

- There isn't a one-to-one mapping between form and function (between the word string and the dialogue act)

  The gun is loaded. *Threat? Warning? Statement?*

- It may require inference (e.g., computing a "conversational implicature"):

  A: Are you going to Paul's party?
  B: I have to work.
  (=> I'm not going — *negative answer*)

NLU

Intent prediction: What is it in practice?

Predict a **meaning representation** given the word string.

In task-oriented dialogue, these are usually “frames” consisting of:

- Domain of the conversation (if not pre-defined)
- Each domain, has a set of possible user intents (task goals).
- Each intent, has a set of possible slots and slot values.

What are possible morning flights from Boston to SF on Tuesday?

```
DOMAIN:        AIR-TRAVEL
INTENT:        SHOW-FLIGHTS
ORIGIN-CITY:   Boston
ORIGIN-DATE:   Tuesday
ORIGIN-TIME:   morning
DEST-CITY:     San Francisco
```

Wake me tomorrow at six.

```
DOMAIN:   ALARM-CLOCK
INTENT:   SET-ALARM
TIME:     2017-07-01 0600-0800
```

NLU

Intent prediction: What is it in practice?

- Many of the NLP techniques you have seen in this course are relevant for intent prediction in dialogue:

  - word embeddings, POS tagging, syntactic parsing, compositional semantics, etc.

- This approach requires **annotated dialogue datasets** where utterances are annotated with meaning representations.

What are possible morning flights from Boston to SF on Tuesday?

```
DOMAIN:        AIR-TRAVEL
INTENT:        SHOW-FLIGHTS
ORIGIN-CITY:   Boston
ORIGIN-DATE:   Tuesday
ORIGIN-TIME:   morning
DEST-CITY:     San Francisco
```

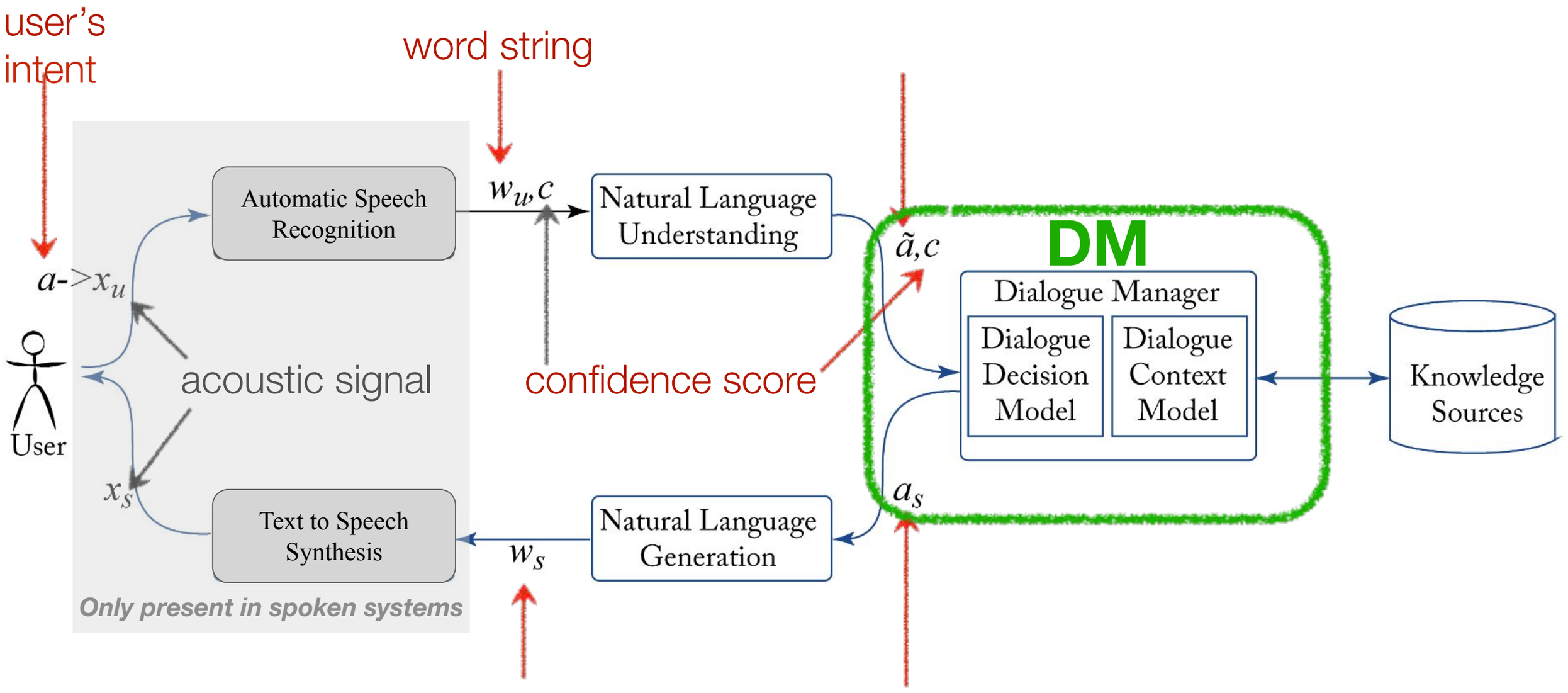
Wake me tomorrow at six.

```
DOMAIN:   ALARM-CLOCK
INTENT:   SET-ALARM
TIME:     2017-07-01 0600-0800
```

Some resources

- Tasks — ParlAI Documentation
- A Survey of Available Corpora for Building Data-Driven Dialogue Systems
- Conversational Dataset List

A dialogue agent (McTear, 2020)



Dialogue management

- The relevant slots may be filled across multiple dialogue turns—the **dialogue context / history** keeps track of this information.
- The **dialogue decision model / policy:** predict the next system action given dialogue context (e.g., slots that are still missing).
  - System intent with the highest probability given the context.

U: Show me morning flights to SF.

```
DOMAIN:        AIR-TRAVEL
INTENT:        SHOW-FLIGHTS
ORIGIN-CITY:   [     ]
ORIGIN-DATE:   [     ]
ORIGIN-TIME:   morning
DEST-CITY:     San Francisco
```

```
DOMAIN:        AIR-TRAVEL
INTENT:        REQUEST(ORIGIN-CITY)
```

S: Where are you flying from?

Dialogue management

Confirmation and rejection

- How likely is the system to have understood the user?
- We can exploit NLU confidence scores to decide on a confirmation/rejection policy:

$< \alpha$	low confidence	reject
$\geq \alpha$	above the threshold	confirm explicitly
$\geq \beta$	high confidence	confirm implictly
$\geq \gamma$	very high confidence	don't confirm at all

`CONFIRM_EXPLICIT(ORIGIN-CITY)`

S:	Which city do you want to leave from?
U:	Baltimore.
S:	**Do you want to leave from Baltimore?**
U:	Yes.

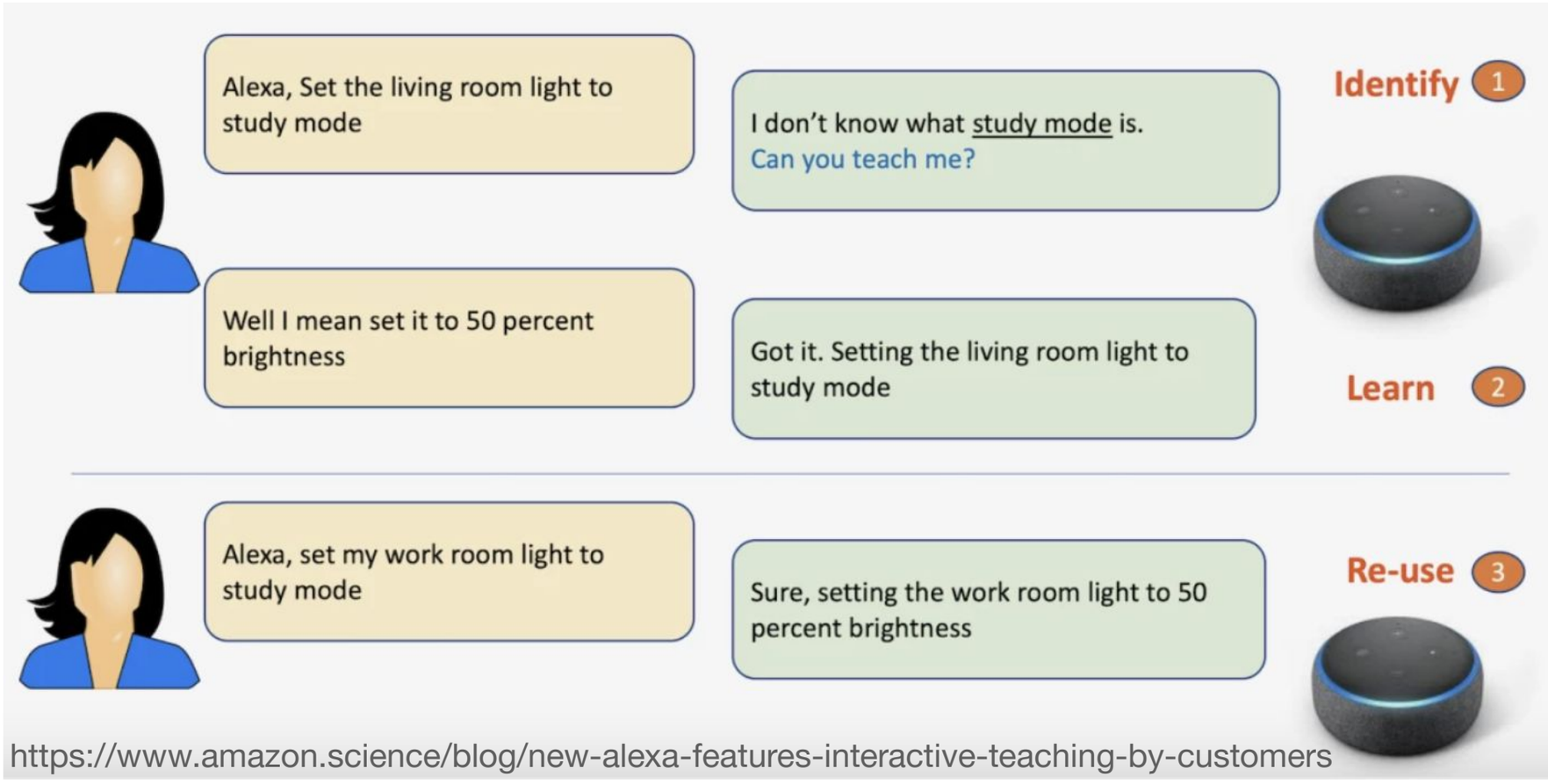
`CONFIRM_IMPLICIT(DEST-CITY)`

U:	I want to travel to Berlin
S:	**When do you want to travel to Berlin?**

Dialogue management

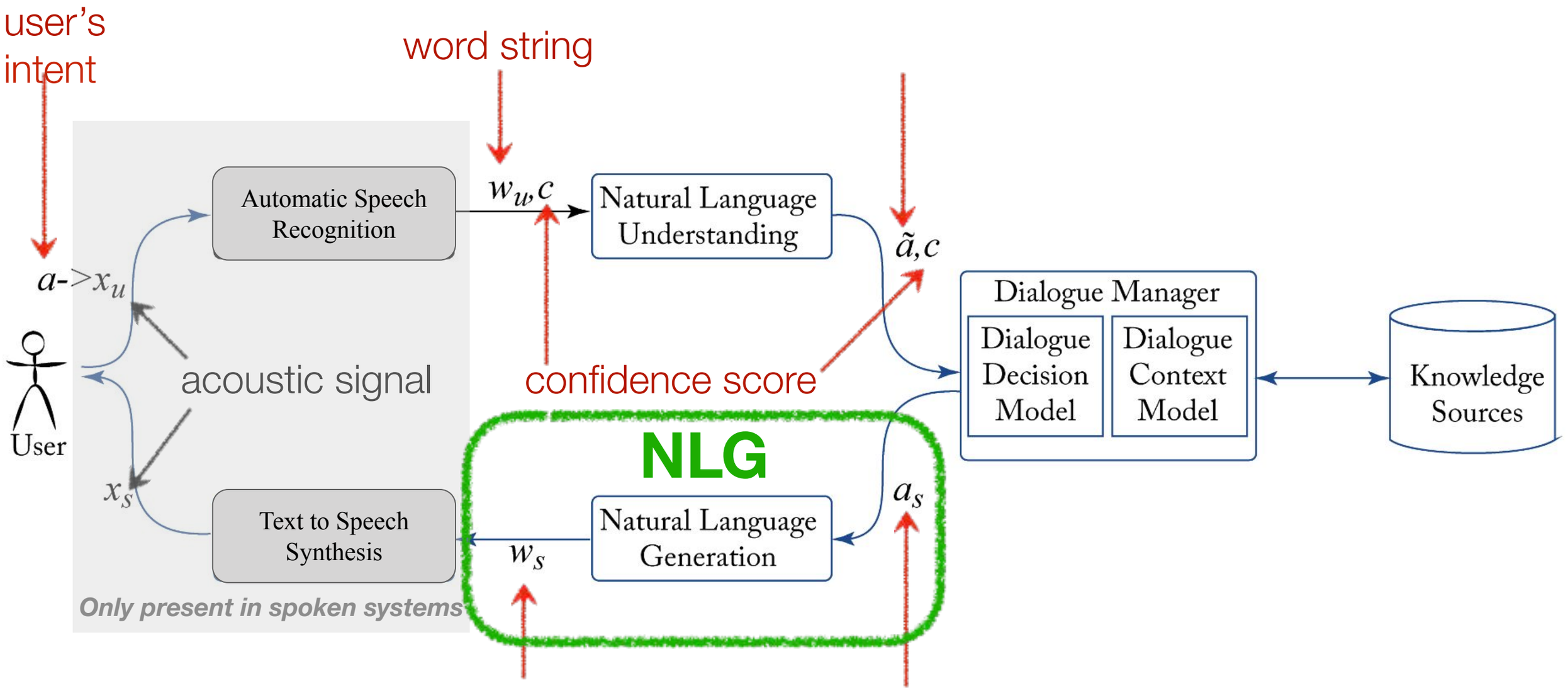
Advanced: Learning and generalisation

- Confidence scores can also be exploited to identify unknown slots and learn to generalise to new situations



https://www.amazon.science/blog/new-alexa-features-interactive-teaching-by-customers

A dialogue agent (McTear, 2020)



NLG

Assuming the DM has chosen a next system action/intent…

- The goal of the NLG module is to learn to generate sentences by training on many representation/sentence pairs from an annotated dialogue corpus

- Some examples:

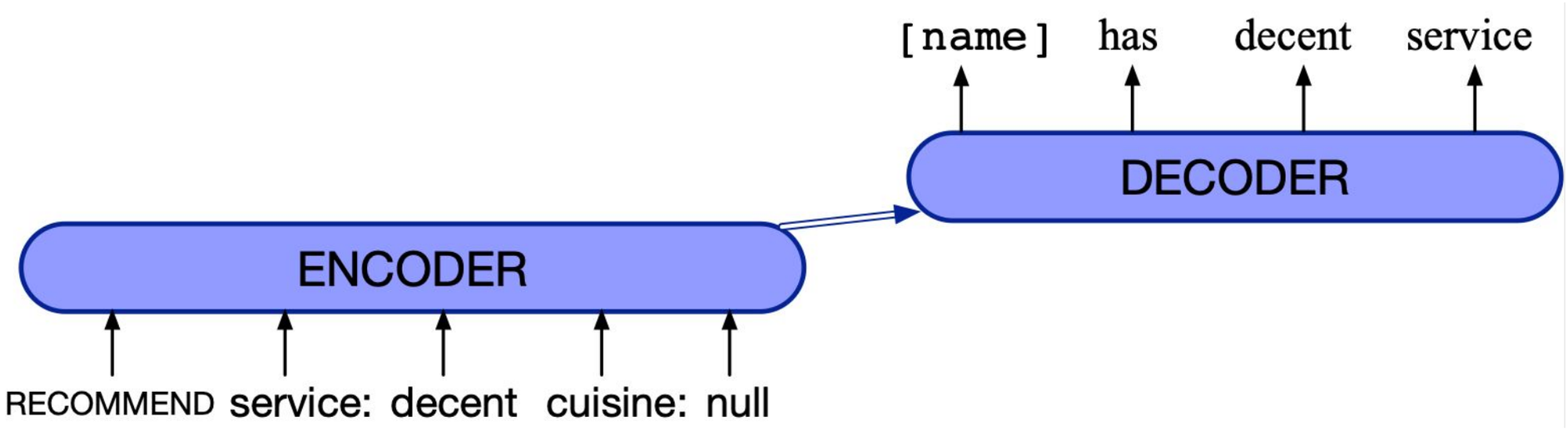
	`recommend(restaurant name= Au Midi, neighborhood = midtown, cuisine = french`
1	Au Midi is in Midtown and serves French food.
2	There is a French restaurant in Midtown called Au Midi.
	`recommend(restaurant name= Loch Fyne, neighborhood = city centre, cuisine = seafood)`
3	Loch Fyne is in the City Center and serves seafood food.
4	There is a seafood restaurant in the City Centre called Loch Fyne.

NLG

Sequence-to-sequence prediction (cf. previous lecture):

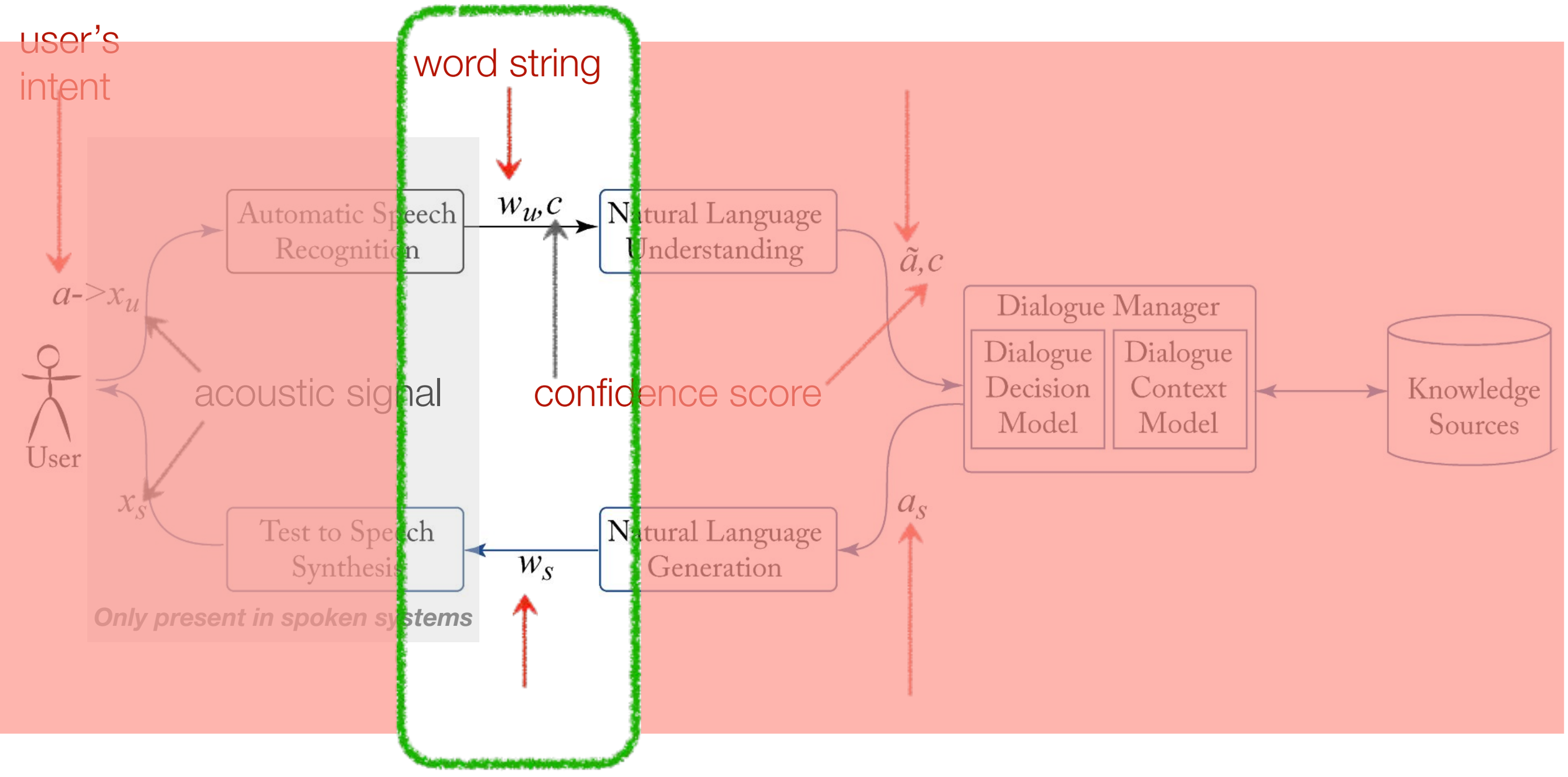
- Input: linearised meaning representation
- Output: word string (system utterance)

	`recommend(restaurant name= Au Midi, neighborhood = midtown, cuisine = french`
1	Au Midi is in Midtown and serves French food.
2	There is a French restaurant in Midtown called Au Midi.
	`recommend(restaurant name= Loch Fyne, neighborhood = city centre, cuisine = seafood)`
3	Loch Fyne is in the City Center and serves seafood food.
4	There is a seafood restaurant in the City Centre called Loch Fyne.



(NB: Delexicalised representation where entities are replaced with general placeholders to help with generalisation)

Non-modular systems



Non-modular systems

Chatbots

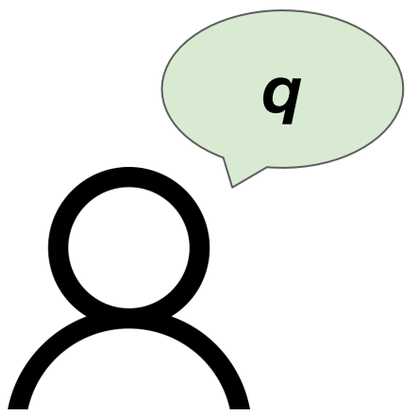
- Dialogue response generation from previous turn(s), without intermediate meaning representations.
- Typically used to model social **chit-chat dialogue** (no need to make progress towards task completion)
- Two methods: Retrieval vs generation

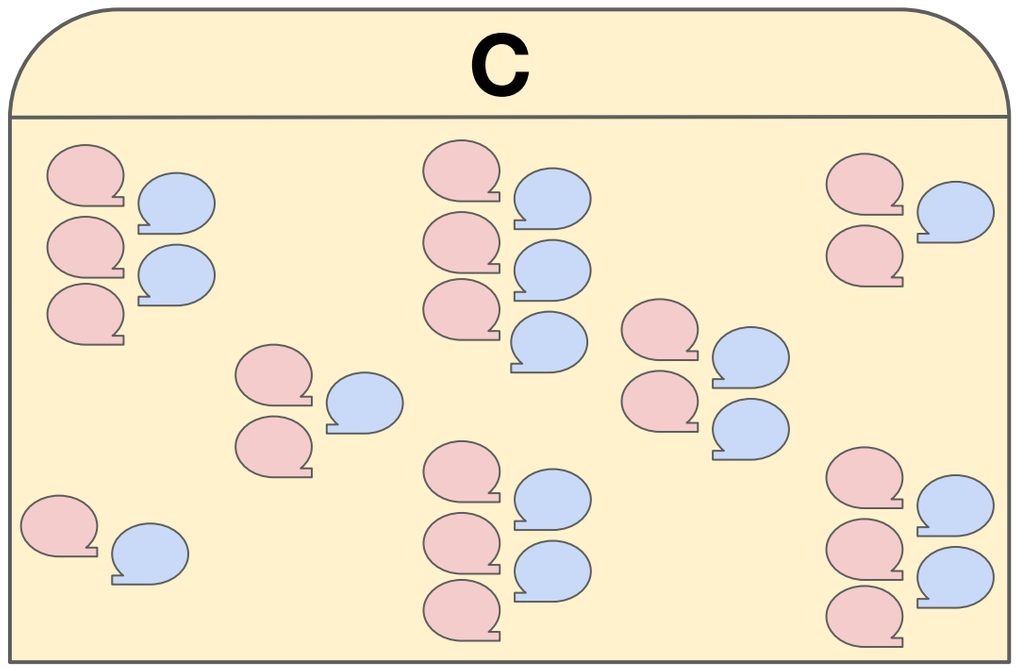
A: What's your favorite holiday?
B: I'm a big fan of Christmas.
A: Is that so? Mine is Holloween.
B: I also like Holloween. But I like Christmas most.

Non-modular systems

Retrieval

- Given a user turn *q* and a dialogue corpus *C*

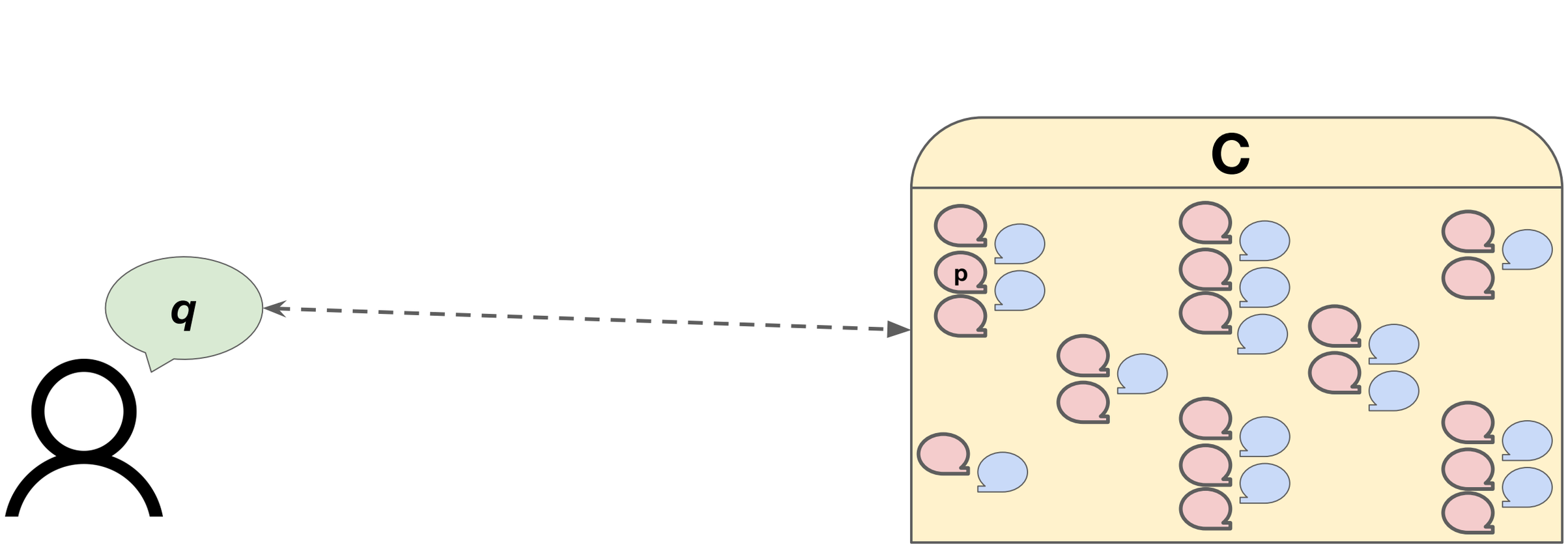




Non-modular systems

Retrieval

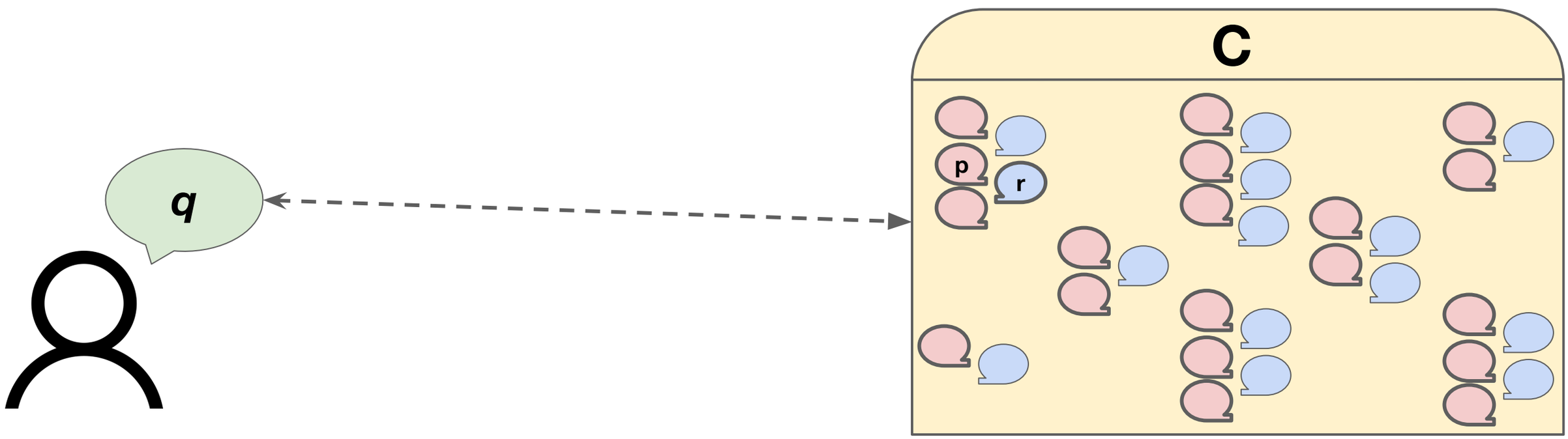
- Given a user turn *q* and a dialogue corpus *C*
- Find in *C* a turn *p* that is most similar to *q*



Non-modular systems

Retrieval

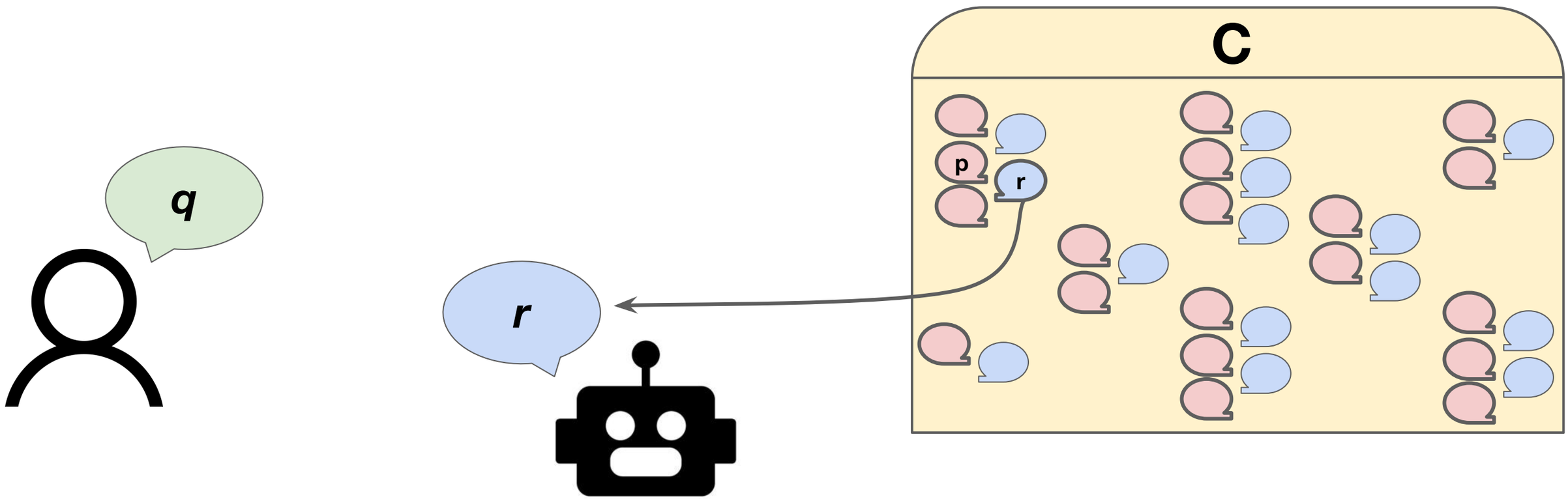
- Given a user turn *q* and a dialogue corpus *C*
- Find in *C* a turn *p* that is most similar to *q*
- Retrieve the turn *r* following *p* in *C*



Non-modular systems

Retrieval

- Given a user turn *q* and a dialogue corpus *C*
- Find in *C* a turn *p* that is most similar to *q*
- Retrieve the turn *r* following *p* in *C*
- Use *r* as a response to *q*

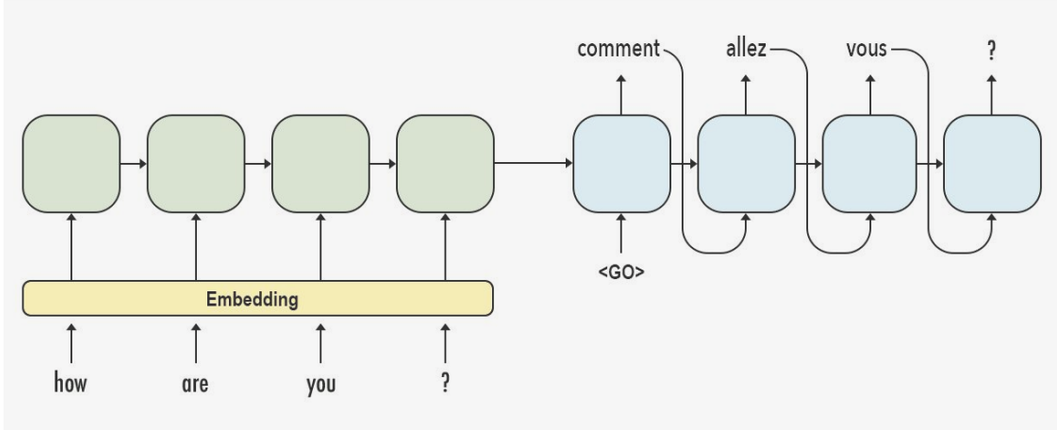


Non-modular systems

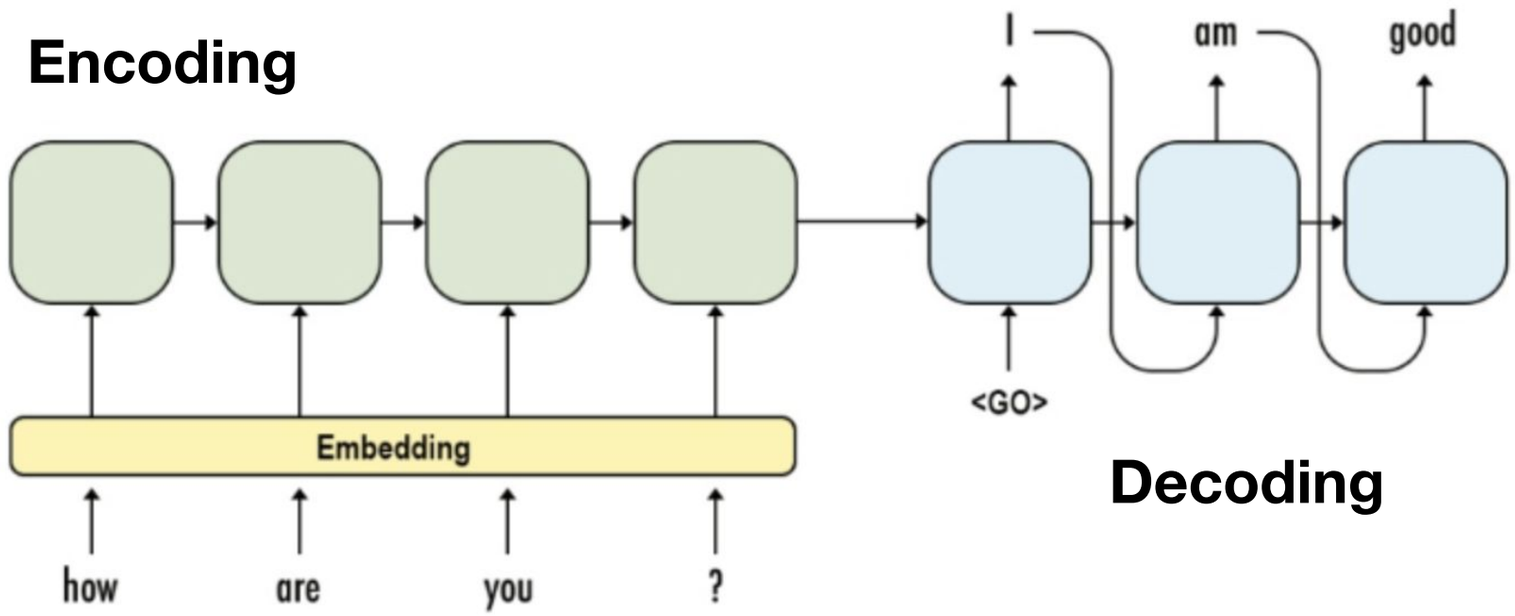
Generation

- **Sequence-to-sequence** models:

  - Inspired by machine translation



- Encoder RNN to produce a representation of the previous turns
- Decoder RNN to generate the response word-by-word by conditioning on the context and the response so far



Non-modular systems

Generation

An alternative to the encoder-decoder architecture:

- Use a large pre-trained **language model** (e.g., GPT-X)
- Fine-tune it on conversational data
- Use the language model directly as a response generator

Non-modular systems

PROS

- No annotations needed
- No finite, predefined set meaning representation

CONS

- Very data-hung: trained on dialogue corpora with hundreds of millions or words
- No real understanding
- Tendency to output generic utterances (uninformative, bland, repetitive)

U:	tell me ... how long have you had this falling sickness ?
S:	i'm not a doctor.
U:	How long does it take you guys to learn the drill?
S:	Like 2 weeks ago!!
U:	Where do you live?
S:	I live in England.
U:	In which city do you live?
S:	I live in Los Angeles.

[adapted from Jurafsky & Martin, SLP-3]

Other important topics

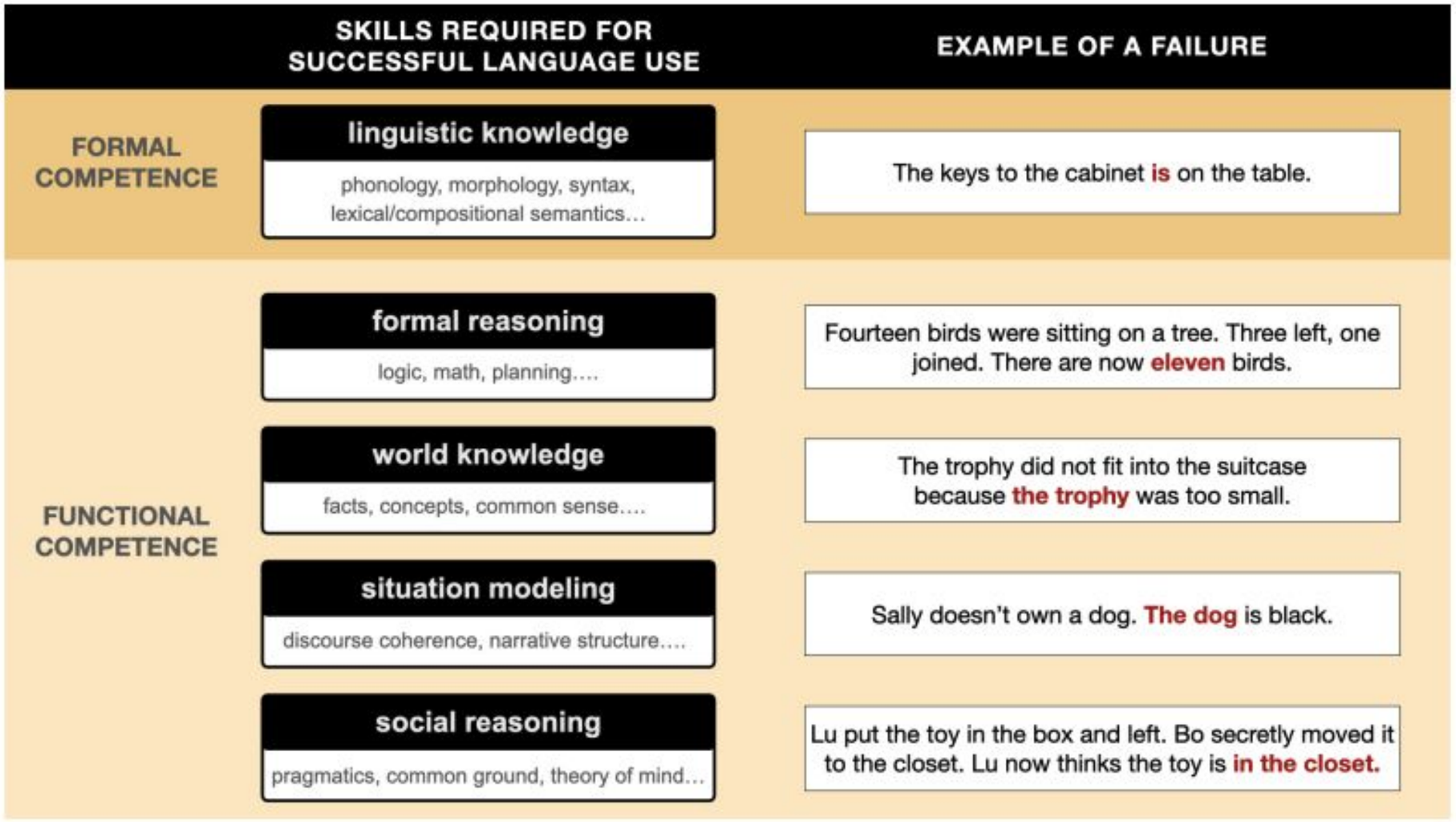
… that we won't be able cover today

- **Evaluation**
  - Very complex and difficult to operationalise
  - Easier for task-oriented dialogue (task completion)

- **Ethical considerations**
  - Implicit biases and reinforcement of stereotypes present in the training data
  - Deception: being perceived as human (anthropomorphism) may be problematic

Shifting to Large Language Models

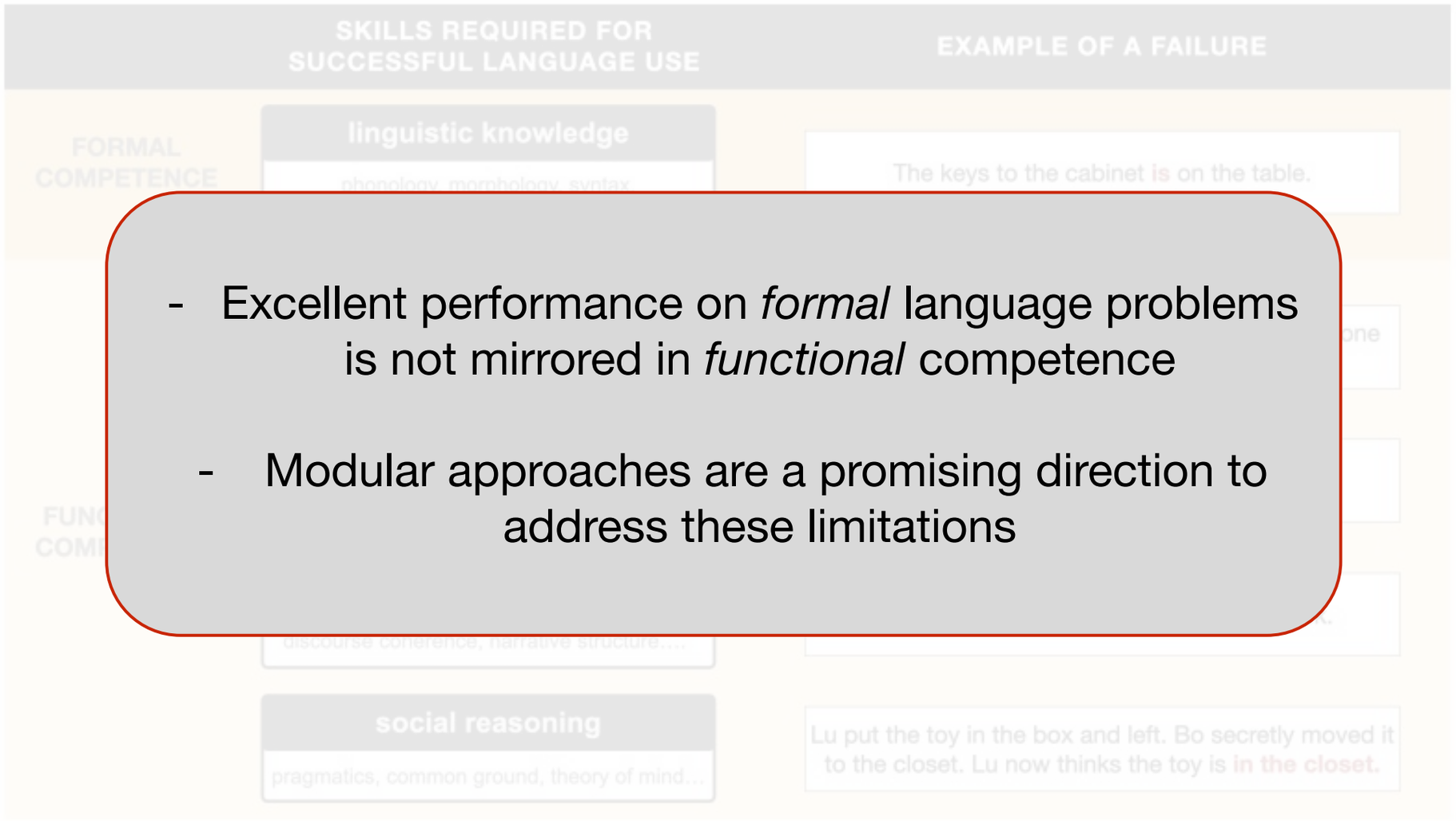
- The field is quickly shifting towards the use of LLMs (see slides from your previous class on 28/11)
- Impressive capabilities to generate well-formed language
- However, their linguistic and cognitive capabilities remain split



Mahowald et al., 2023

Shifting to Large Language Models

- The field is quickly shifting towards the use of LLMs (see slides from your previous class on 28/11)
- Impressive capabilities to generate well-formed language
- However, their linguistic and cognitive capabilities remain split



- Excellent performance on *formal* language problems is not mirrored in *functional* competence
- Modular approaches are a promising direction to address these limitations

Mahowald et al., 2023

Vision & Language Dialogue Systems

VQA

Q: How many people on wheelchairs ?

A: Two

Q: How many wheelchairs ?

A: One

Captioning

Two people are in a wheelchair and one is holding a racket.

Visual Dialog

Q: How many people are on wheelchairs ?

A: Two

Q: What are their genders ?

A: One male and one female

Q: Which one is holding a racket ?

A: The woman

Visual Dialog

Q: What is the gender of the one in the white shirt ?

A: She is a woman

Q: What is she doing ?

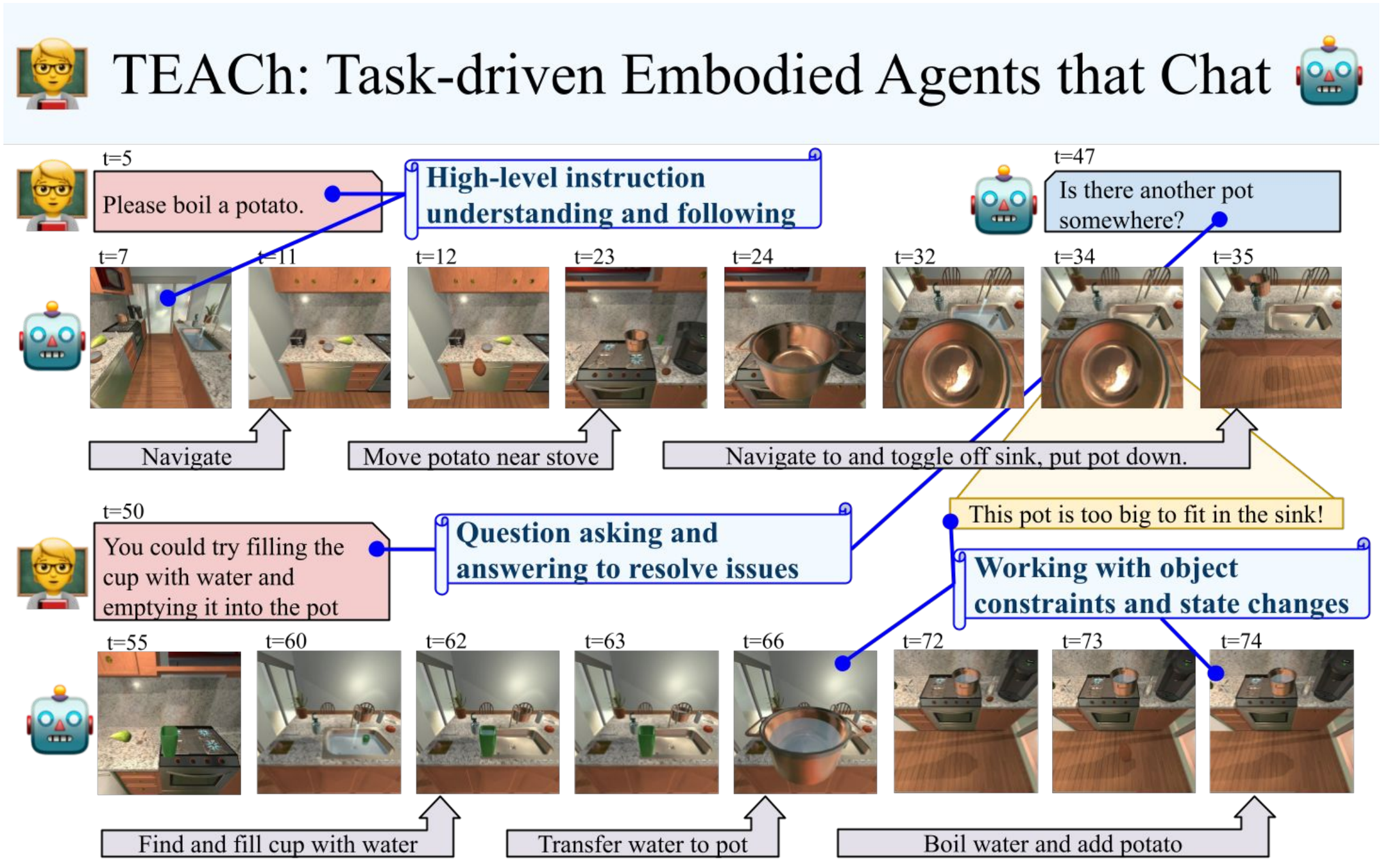
A: Playing a Wii game

Q: Is that a man to her right

A: No, it's a woman

Das et al., CVPR 2017

Vision & Language Dialogue Systems



Padmakumar et al., AAAI 2022

Vision & Language Dialogue Systems

Vizwiz grand challenge: Answering visual questions from blind people

Q: Does this foundation have any sunscreen?
A: yes

Q: What is this?
A: 10 euros

Q: What color is this?
A: green

Q: Please can you tell me what this item is?
A: butternut squash red pepper soup

Q: What type of pills are these?
A: unsuitable image

Q: What type of soup is this?
A: unsuitable image

Q: Who is this mail for?
A: unanswerable

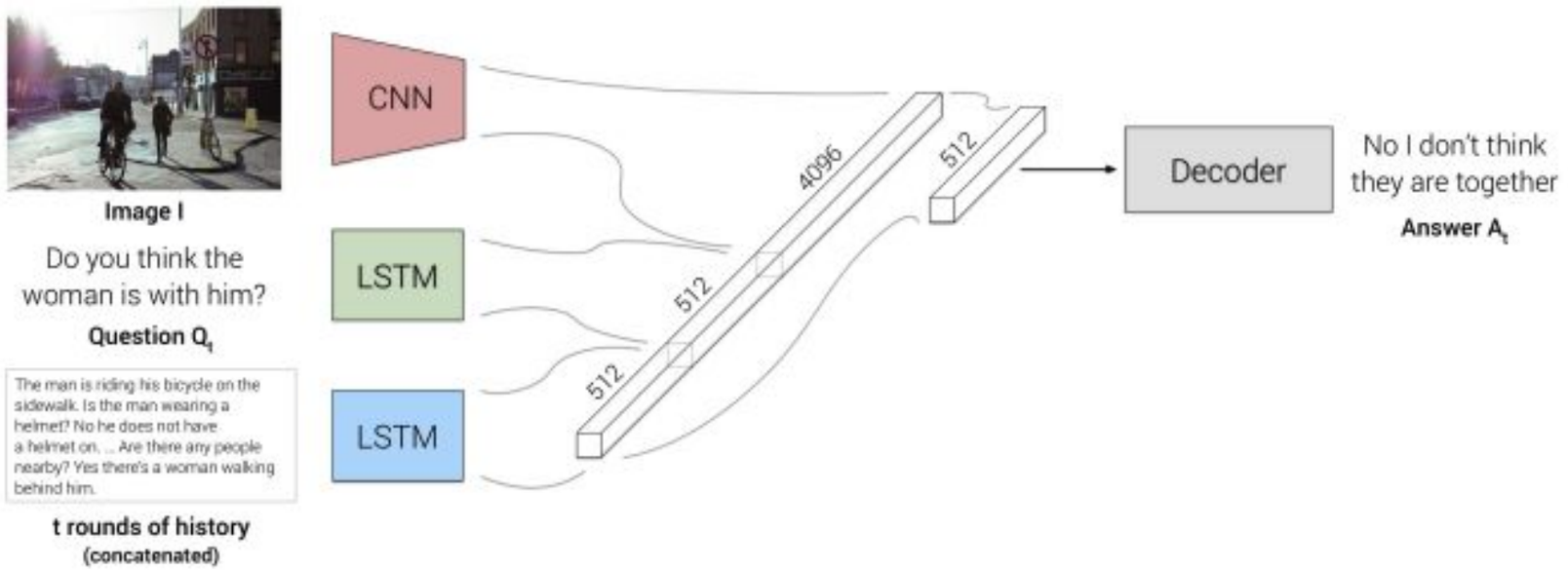
Q: When is the expiration date?
A: unanswerable

Gurari et al., CVPR 2018

Vision & Language Dialogue Systems

Some of the Key Challenges of V&L dialogue systems

- Integrating different modalities



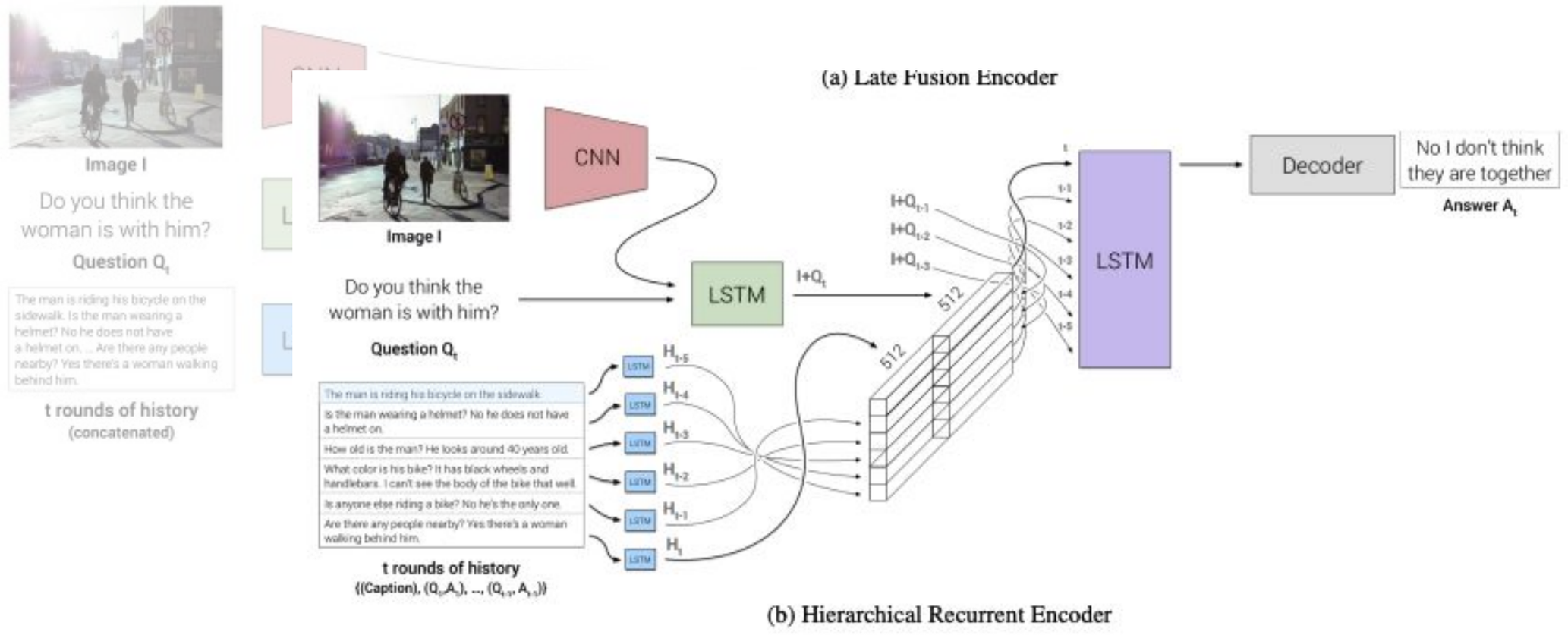
(a) Late Fusion Encoder

Das et al., CVPR 2017

Vision & Language Dialogue Systems

Some of the Key Challenges of V&L dialogue systems

- Integrating different modalities

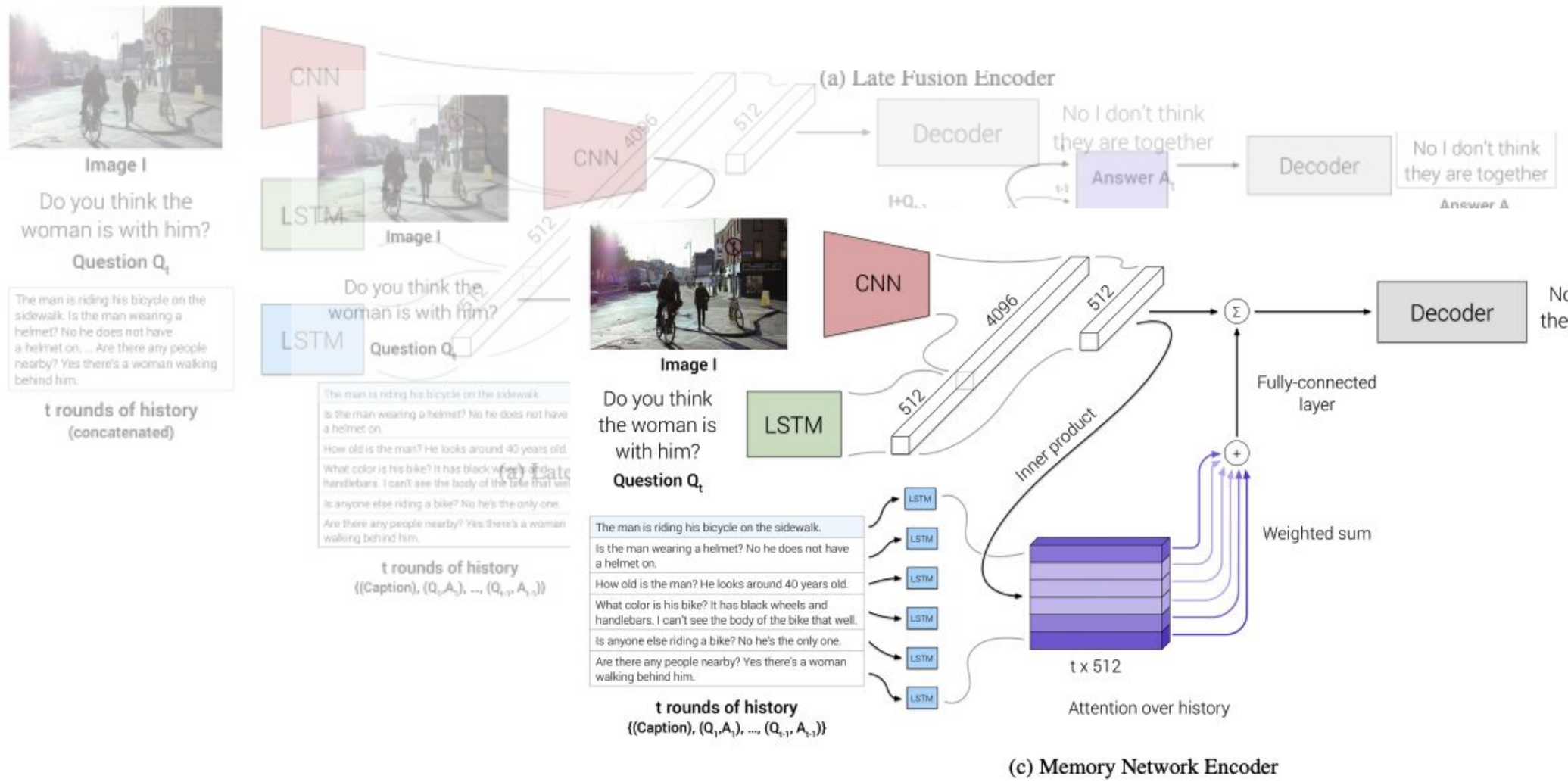


Das et al., CVPR 2017

Vision & Language Dialogue Systems

Some of the Key Challenges of V&L dialogue systems

- Integrating different modalities

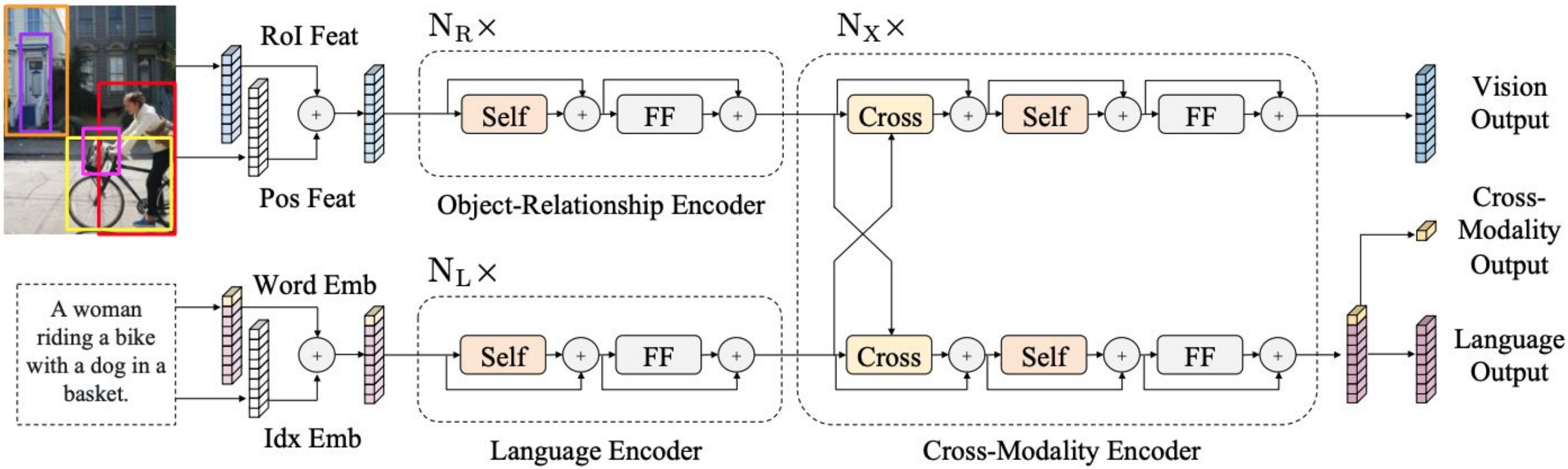


Das et al., CVPR 2017

Vision & Language Dialogue Systems

Some of the Key Challenges of V&L dialogue systems

- Integrating different modalities



Tan and Bansal, EMNLP 2019

Vision & Language Dialogue Systems

Some of the Key Challenges of V&L dialogue systems

- Generating dialogue exchanges coherent with the visual input: hallucinations

"I've Seen Things You People Wouldn't Believe":
Hallucinating Entities in GuessWhat?!

Alberto Testoni
DISI, University of Trento
Trento, Italy
alberto.testoni@unitn.it

Raffaella Bernardi
CIMeC, DISI, University of Trento
Rovereto, Italy
raffaella.bernardi@unitn.it

Abstract

Natural language generation systems have witnessed important progress in the last years, but they are shown to generate tokens that are unrelated to the source input. This problem affects computational models in many NLP tasks, and it is particularly unpleasant in multimodal systems. In this work, we assess the rate of object hallucination in multimodal conversational agents playing the GuessWhat?! referential game. Better visual processing has been shown to mitigate this issue in image captioning; hence, we adapt to the GuessWhat?! task the best visual processing models at dis-

is it a dog ? no
is it a chair ? no
is it a fridge ? no
is it a cup ? yes
on the right? yes

is it a person ? no
is it a skateboard ? no
is it a car ? yes
is it white ? no
is it green ? no

Testoni and Bernardi, ACL-SRW 2021

Vision & Language Dialogue Systems

Some of the Key Challenges of V&L dialogue systems

- Planning and Reasoning skills to ask strategically informative questions

Looking for Confirmations:
An Effective and Human-Like Visual Dialogue Strategy

Alberto Testoni
DISI - University of Trento
Trento - Italy
alberto.testoni@unitn.it

Raffaella Bernardi
CIMeC and DISI - University of Trento
Rovereto (TN) - Italy
raffaella.bernardi@unitn.it

Abstract

Generating goal-oriented questions in Visual Dialogue tasks is a challenging and long-standing problem. State-Of-The-Art systems are shown to generate questions that, although grammatically correct, often lack an effective strategy and sound unnatural to humans. Inspired by the cognitive literature on information search and cross-situational word learning, we design *Confirm-it*, a model based on a beam search re-ranking algorithm that guides an effective goal-oriented strategy by asking questions that confirm the model's conjecture about the referent. We take the GuessWhat?! game as a case-study. We show that dialogues generated by *Confirm-it* are more natural and

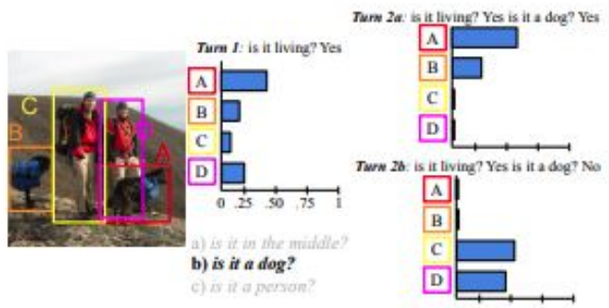


Figure 1: At turn 2, among the questions proposed by the beam search, *Confirm-it* chooses *b* since it is the most suitable one to confirm the current conjecture.

jects, they tend to be overspecific and prefer properties irrespectively of their utility for identifying

Algorithm 1 The *Confirm-it* algorithm

Require: History H, Beam size B, Max turns T,
Image I, Distractors $D_{1:N-1}$, target o_t,
Require: Candidates $C_{1:N} \leftarrow D_{1:N-1} + o_t$
Require: Internal Oracle IO
Require: Target-aware external $Oracle$
for $turn = 1 : T$ **do**
  $\hat{p}(c_{k_{1:N}}) \leftarrow Guesser(H, I, C_{1:N})$
  $c_h \leftarrow argmax(\hat{p}(c_{k_{1:N}}))$
  $q_{1:B} \leftarrow QGen(H, I)$
  $a^*_{1:B} \leftarrow IO(H + q_{1:B}, c_h)$
  $H^*_{1:B} = H + (q_{1:B}, a^*_{1:B})$
  $p^* \leftarrow Guesser(H^*_{1:B}, I, C_{1:N})$
  $Q \leftarrow q_{1:B}[argmax(p^*(c_h))]$
  $Oracle$ provides an answer A to Q
  $H \leftarrow H + Q + A$
end for

Testoni and Bernardi, EMNLP 2021

End of Part 1

Gestures In Face-to-face Dialogues

Detection & Alignment with Speech

Understanding and Modeling **Multimodal Alignment** In Face-to-Face Dialogue

- Project: Gestures
  - What are gestures, and why is it interesting to detect them automatically?
  - Gestural alignment

- Computational models for detecting gestures
  - Skeletal models
  - Speech and Skeletal models

Understanding and Modeling **Multimodal Alignment** In Face-to-Face Dialogue

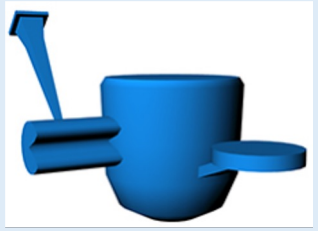
- Multimodal alignment
  - Gesturing and speaking → building common ground (conceptual pacts) → alignment → mutual understanding
- What is the role of co-speech gestures in the collaborative process of creating a mutual understanding of referring expressions?

- Face-to-face dialogue
  - The most common way of communication!

Instances of Speech and Gestural Alignment

- **Round 1 (A-D):** en aan de andere kant heeft hij een soort van raarvormige neus met een **toeter** eraan kan je zeggen
  *and on the other side it has a kind of nose with a strange form with a **horn** so to say*
- **Round 2 (A-M):** ja zo'n **toeter**
  *yes kind of **horn***
- **Round 3 (A-D):** uh deze die houdt de weer zo'n dienblad vast met de **toeter**
  *uh, this one holds again this kind of tray with the **horn***
- **Round 3 (A-D):** dienblad en toeter #laughs#
  *tray and **horn** #laughs#*
- **Round 4 (B-D):** dit is uh die het dienblad vasthoudt en de **toeter**
  *this is uh the one holding the tray and **horn***
- Round 5 (A-D): uh deze heeft de dienblad en de **toeter**
  *uh this one has the tray and the **horn** #laughs#*
- Round 6 (A-M): **toeter**
  horn
- Round 6 (B-D): **toeter**
  horn

Pair	Expression	Sequence of speakers and rounds
Pair 9	toeter	[A, A, A, A, B, A, A, B] [1, 2, 3, 3, 4, 5, 6, 6]



Gestures

- Why do we gesture?
- Gestures that are not necessarily co-speech
  - Head Gestures, Facial Expressions, etc
  - Emblematic gestures
    - Conventionalized meaning & culture-language specific
    - Can also be independent
- Co-speech gestures
  - Different from other aspects of nonverbal behavior due to their tight link with speech: semantically, pragmatically, and temporally
  - Along with speech: they constitute the human language.

Classifying Co-speech Gestures

- Representational: eco or elaborate the meaning of co-occurring speech
  - Iconic
    - E.g., performing action or portraying shape
  - Metaphoric
    - Iconic gestures portraying abstract content
- Non-representational:
  - Deictic gesture → pointing Gestures
  - Beat gestures
    - Short, repetitive movement → correlates with speech prosody
    - Co-occur always with speech → no semantic information
    - E.g., emphasize certain parts of speech

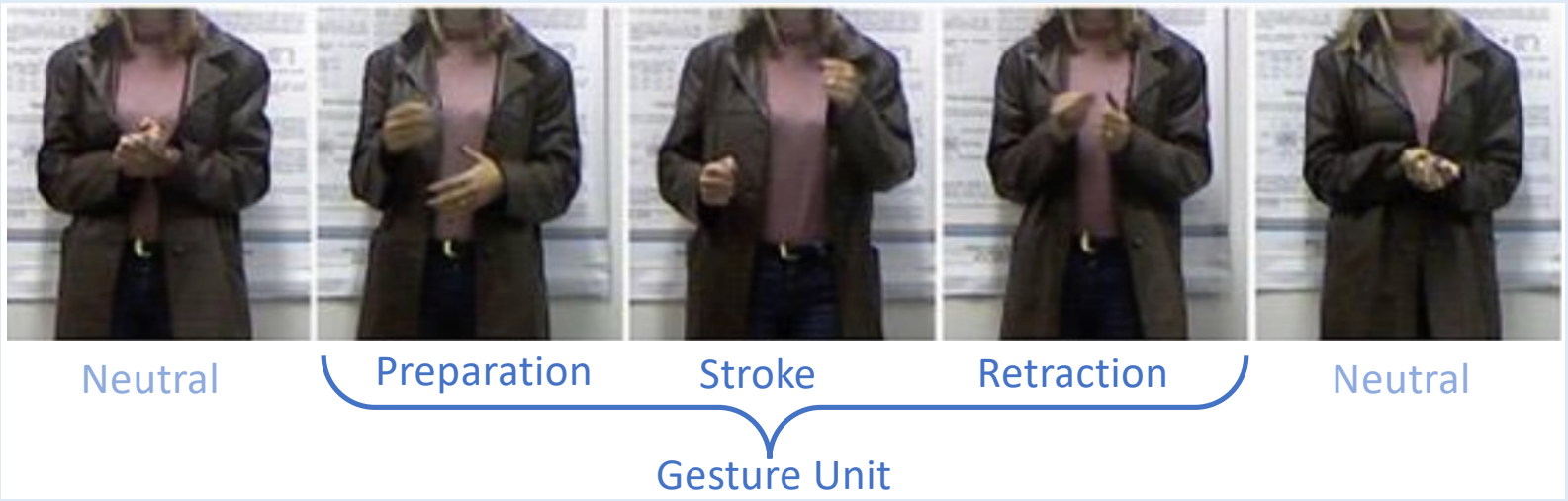
Co-Speech Gesture Detection through Multi-Phase Sequence Labeling

Esam Ghaleb[A], Ilya Burenko[B], Marlou Rasenberg[C], Wim Pouw[D], Peter Uhrig[B], Judith Holler[D], Ivan Toni[D], Aslı Özyürek[D,E], Raquel Fernández[A]

A University of Amsterdam (UvA)
B ScaDS.AI Dresden/TU Dresden
C The Meertens Institute, the Royal Netherlands Academy of Arts and Sciences (KNAW)
D Radboud University (RU)
E Max Planck Institute for Psycholinguistics (MPI)

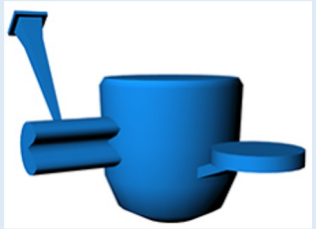
Limitations of Current Approaches & Our Novelty

- Limitations:
  - Silent gestures & limited number of gestures
  - Binary approach
- Gestures unfold over time, often following predictable movement phases
- Novelty: Multi-phase co-speech gesture detection
  - They are linked with speech: semantically, pragmatically, and temporally
  - We focus on detecting co-speech gestures in naturalistic, conversational data



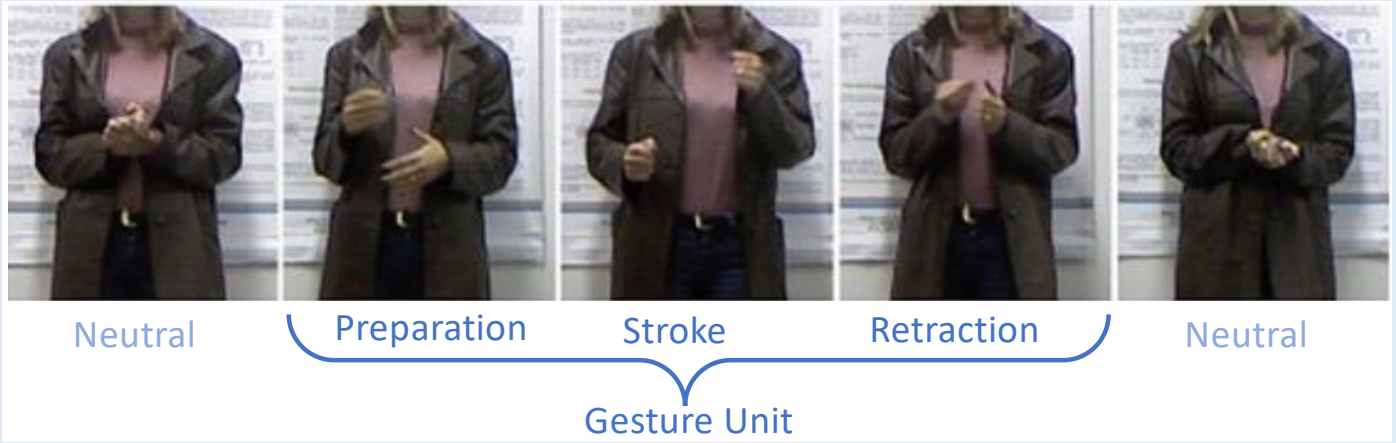
Dataset (Rasenberg et al., 2022)

- 19 face-to-face task-oriented dialogues
  - 38 subjects with 16 hours of recordings
- Referential game
  - One participant describes a novel object while the other participant tries to find it among 16 candidates, using any speech and **gestures**
- Each gesture stroke was manually segmented:
  - 6106 gestural strokes with an average duration of 0.58 seconds

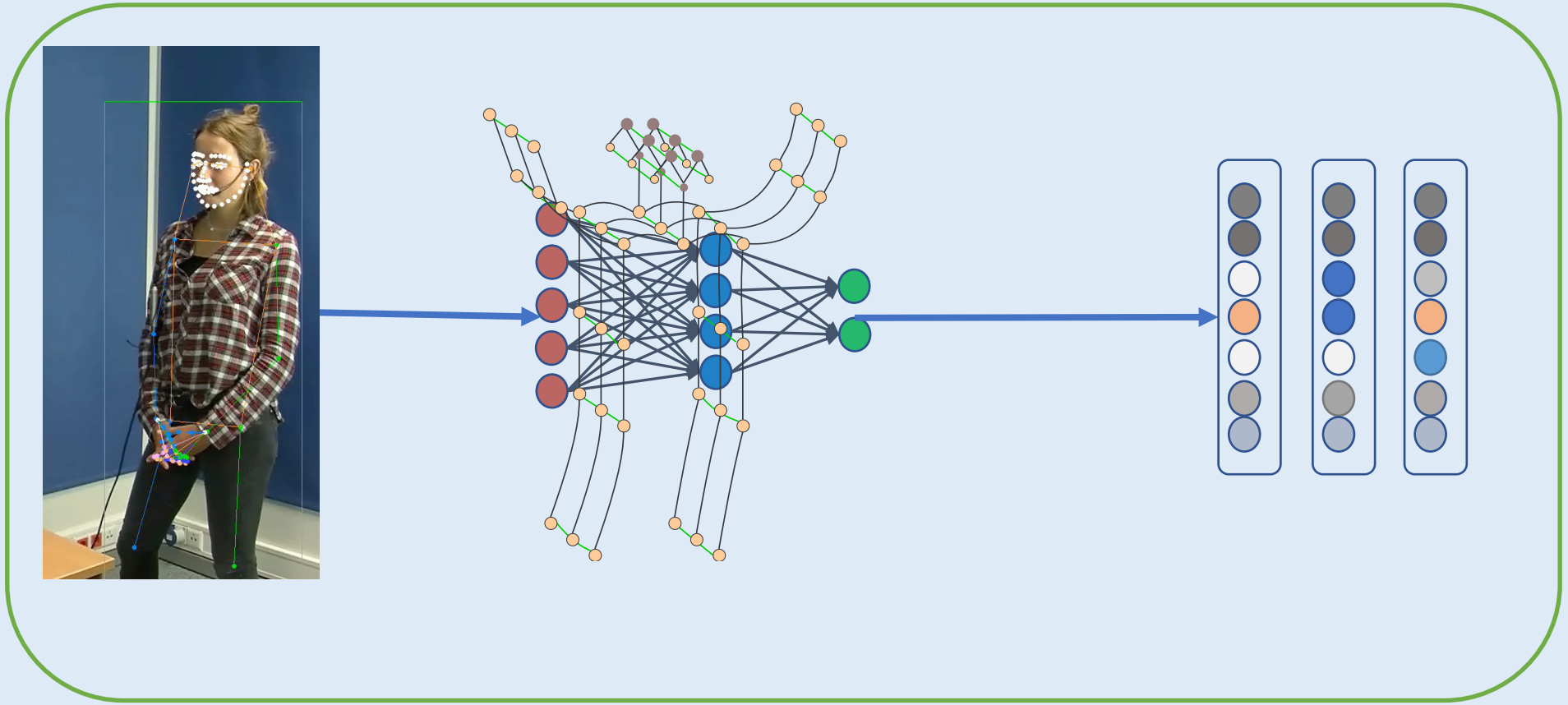


Constructing Multi-Phase Sequential Data: Sliding Window

- **Time Window Duration:** 0.58 seconds
- **Window Shift:** offset by 2 frames for each shift
- **Labeling criteria**
  - **Preparation Phase**
  - **Stroke Phase**
  - **Retraction Phase**
  - **Neutral Phase**



Input Data & Embedding Model: Skeleton-based Gesture Detection



Input videos

Detection model: Graph Convolutional Networks (GCNs)

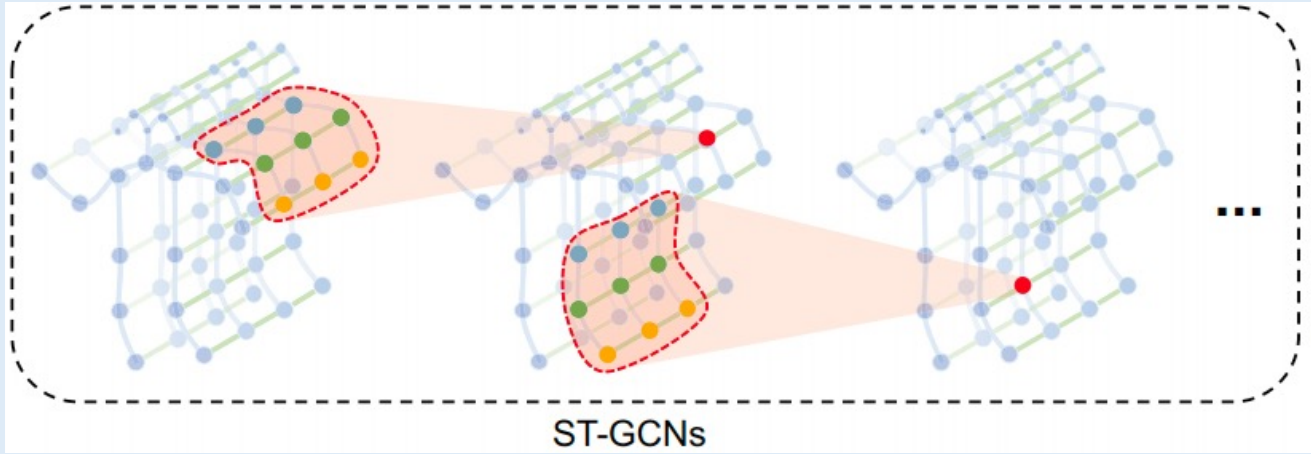
Gesture phase detection

Spatial-Temporal Graph Convolutional Network (Yan et al. 2018)

- Construct a spatial-temporal graph on skeleton sequences

- Apply multiple layers of spatial-temporal graph convolution (ST-GCN) on the graph to gradually generate high-level features

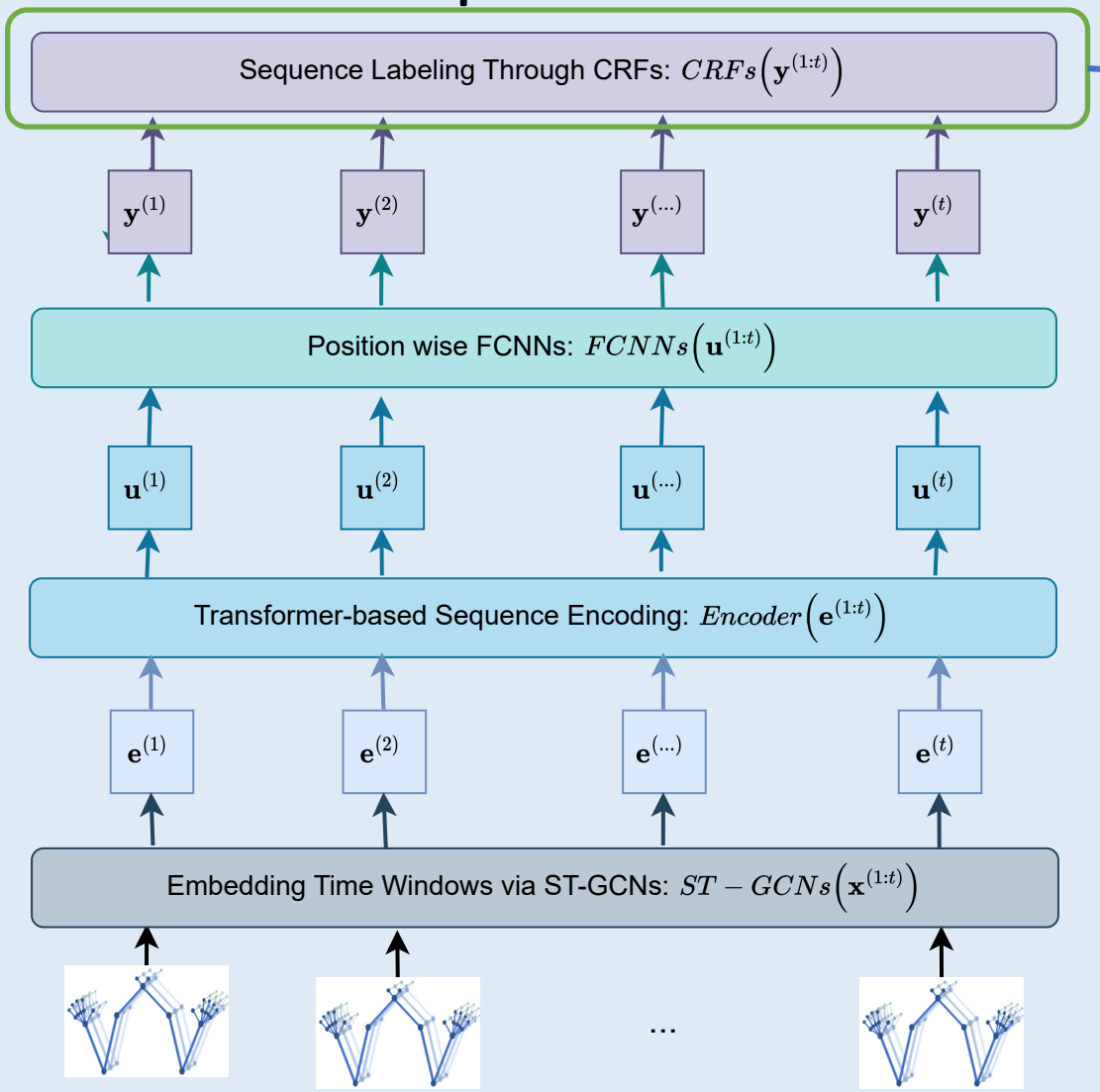
Pose Estimation

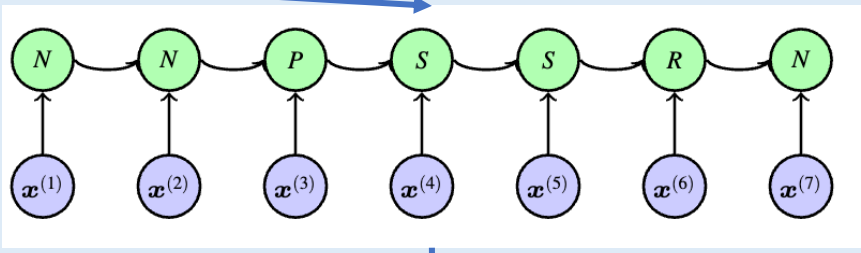


Gesture segmentation

E.g., gestures

Gesture Detection through Multi-Phase Sequence Labelling



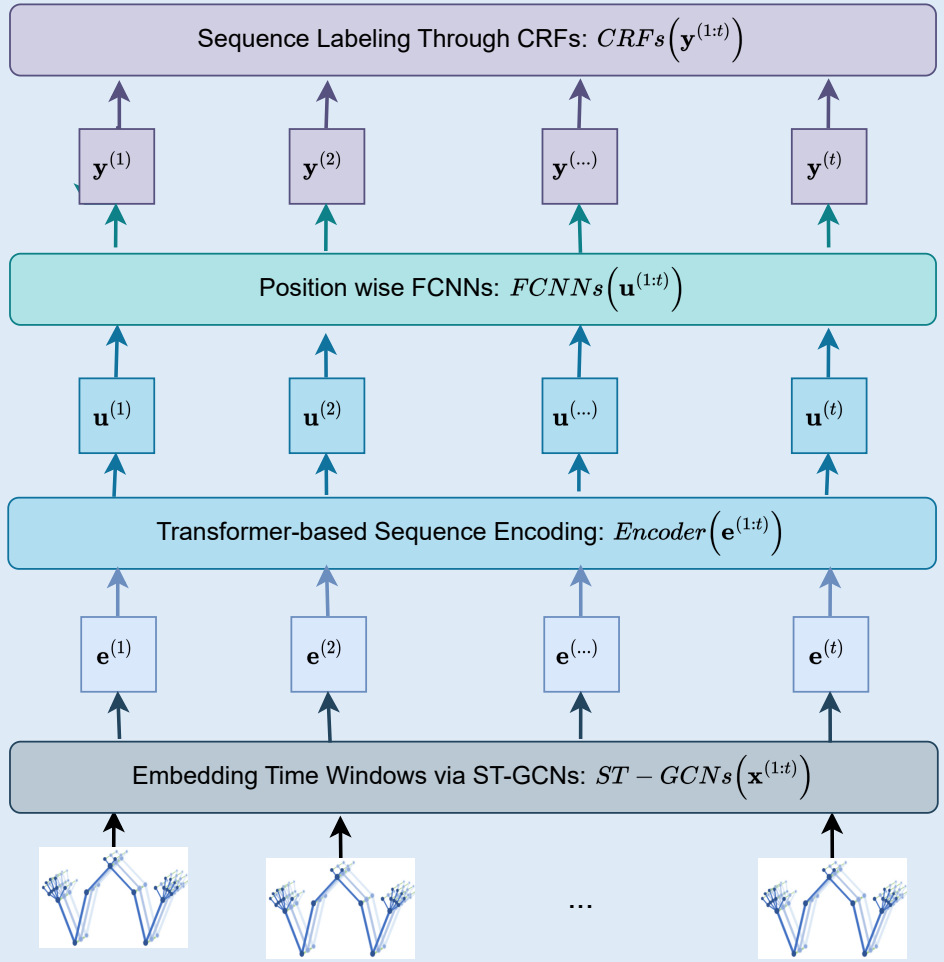


$$P(\boldsymbol{y}^{(1:t)}|\boldsymbol{x}^{(1:t)}) = \prod_{i-1}^{t} P(\boldsymbol{y}^{(i)}|\boldsymbol{x}^{(1:t)}, \boldsymbol{y}^{(i-1)})$$

Ablation Study & Models Comparisons

- **Sequential** vs. **classification approach**
  - **Sequence labeling via CRFs**
  - **Classification approach:** instead of applying sequence labeling via CRFs, phases are classified independently

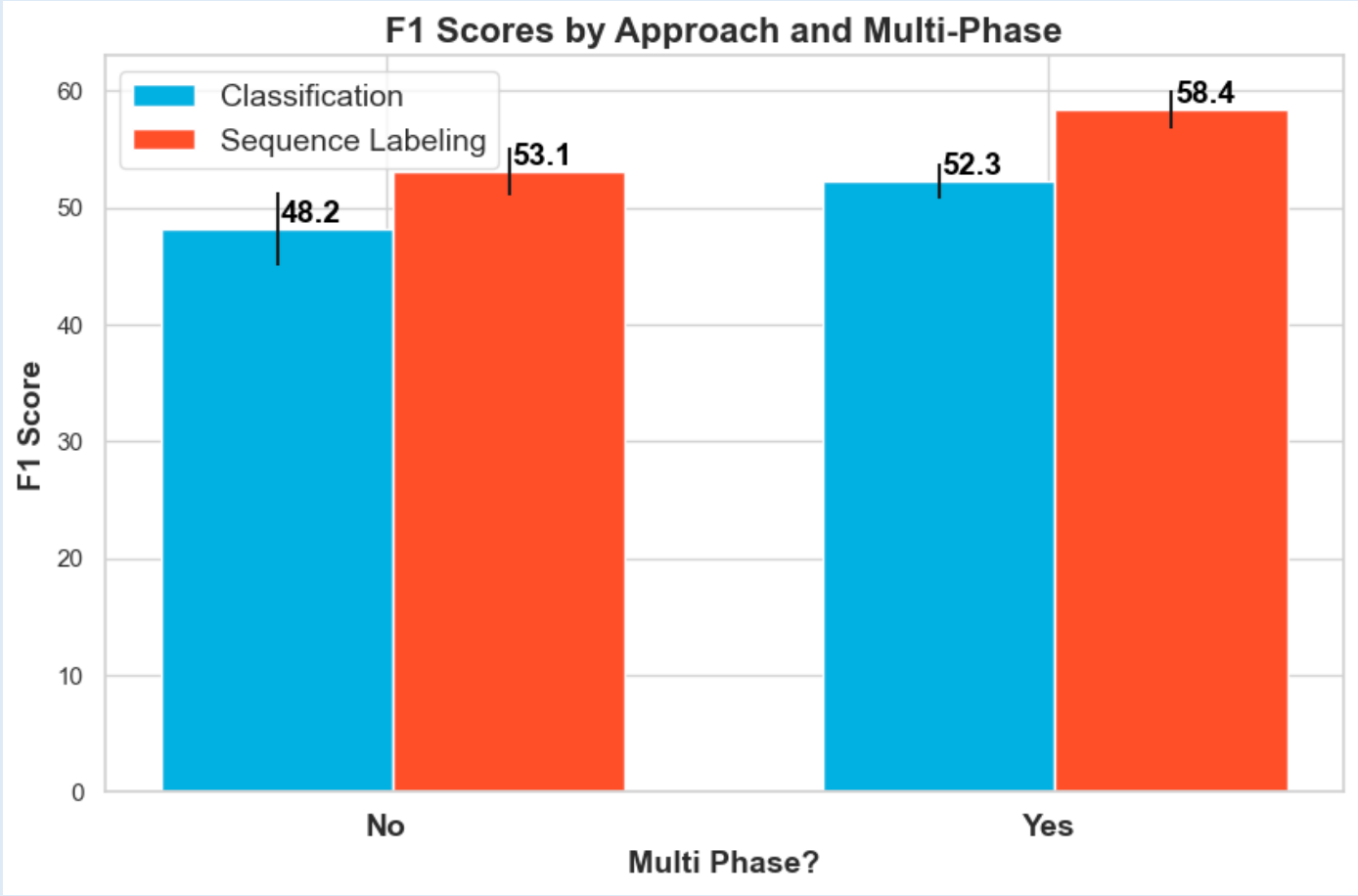
- **Multi-phase** vs. **binary approach**
  - Multi-phase approach: preparation, stroke, retraction, and neutral phases
  - The binary approach simplifies the labeling process by focusing on stroke detection: stroke vs neutral



Gesture Detection through Multi-Phase Sequence Labelling

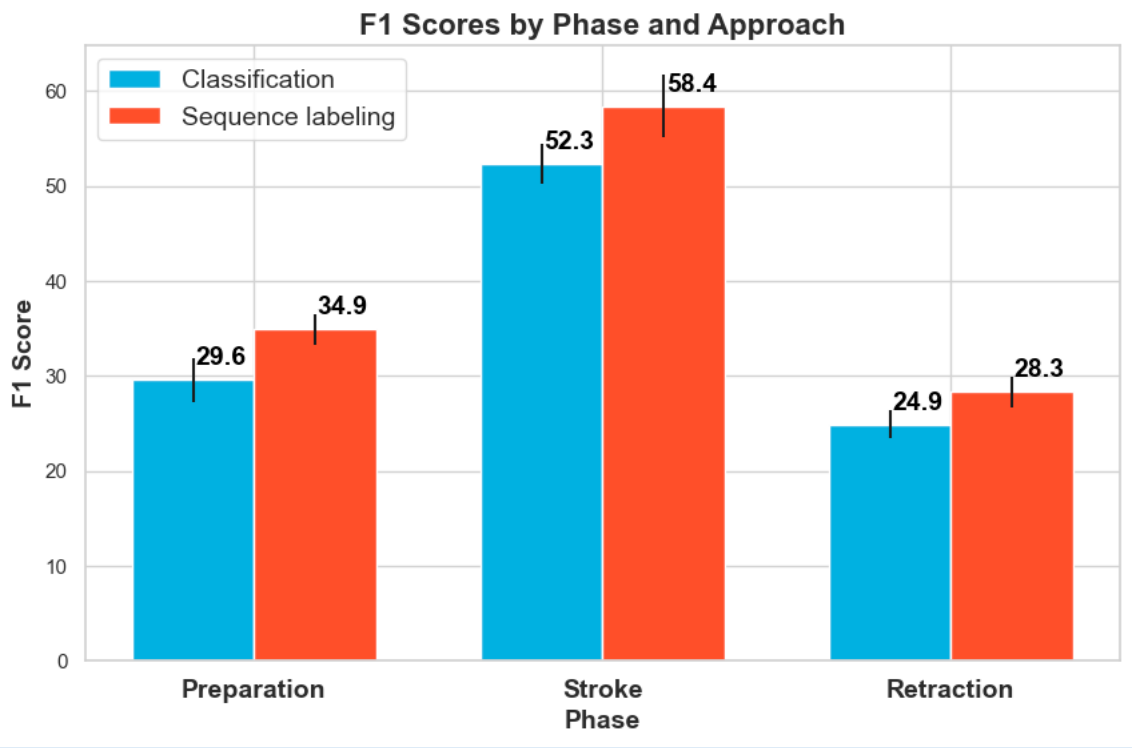
- Conceptualizing gesture detection as sequence labelling gives better performance than a classification approach

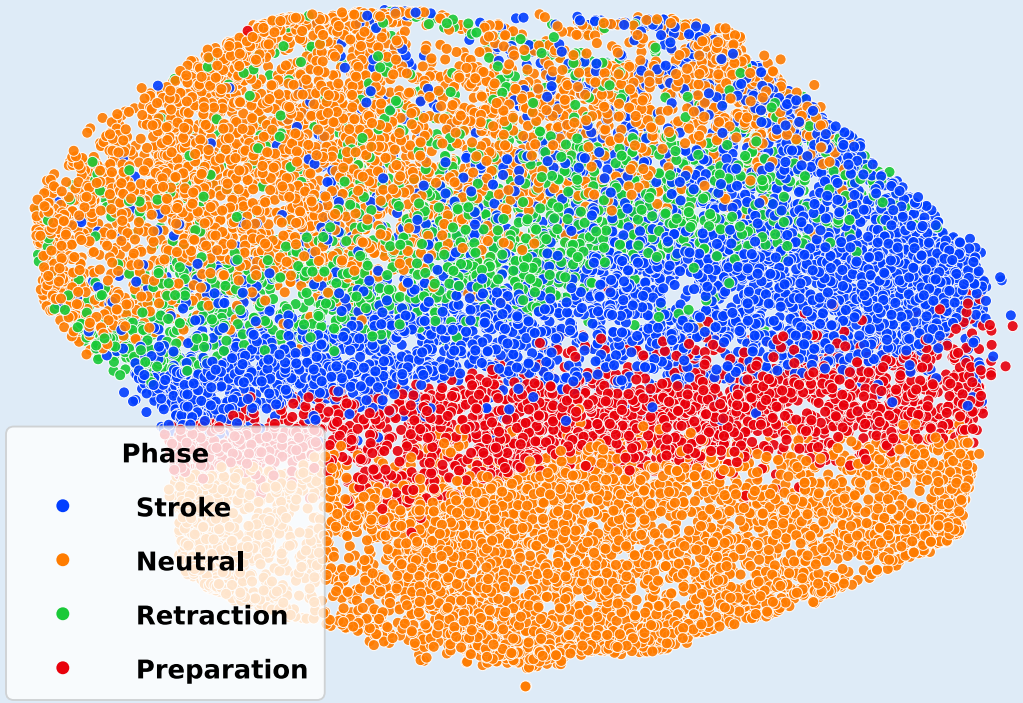
- Multi-phase (labeling or classification) is better than the binary approach



Analysis: Performance on Gesture Phases

- We observed that all models are better at detecting the gesture stroke than its boundaries, particularly the retraction phase





Conclusions

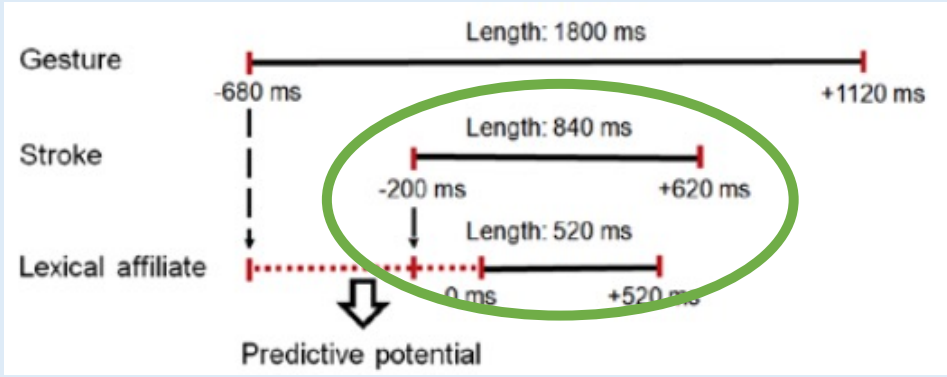
- We proposed a novel framework that emphasizes the structured and sequential nature of gestures:
  - Focusing on co-speech gestures in naturalistic, conversational data.
- Our framework reframes the gesture detection task as a multi-phased sequence labeling problem
- The results show that sequence labeling methods outperform classification approaches in gesture stroke detection

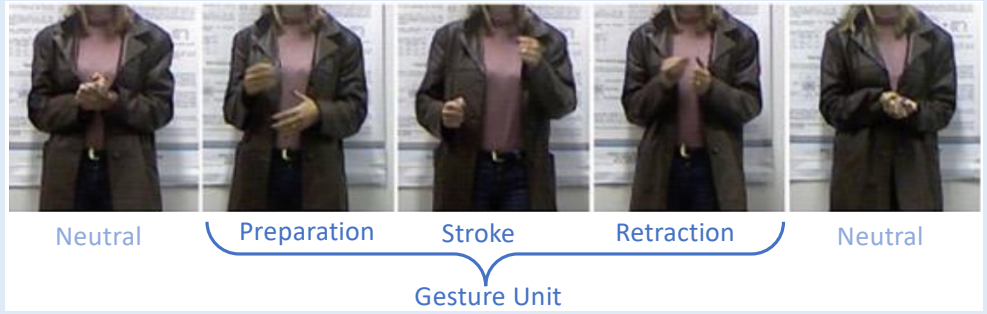
Alignment of Speech and Skeletal Models for Gesture Detection

Esam Ghaleb, Ilya Burenko, Marlou Rasenberg, Wim Pouw, Peter Uhrig, Judith Holler, Ivan Toni, Aslı Özyürek, Raquel Fernández

Temporal Coordination of Speech and Gestures

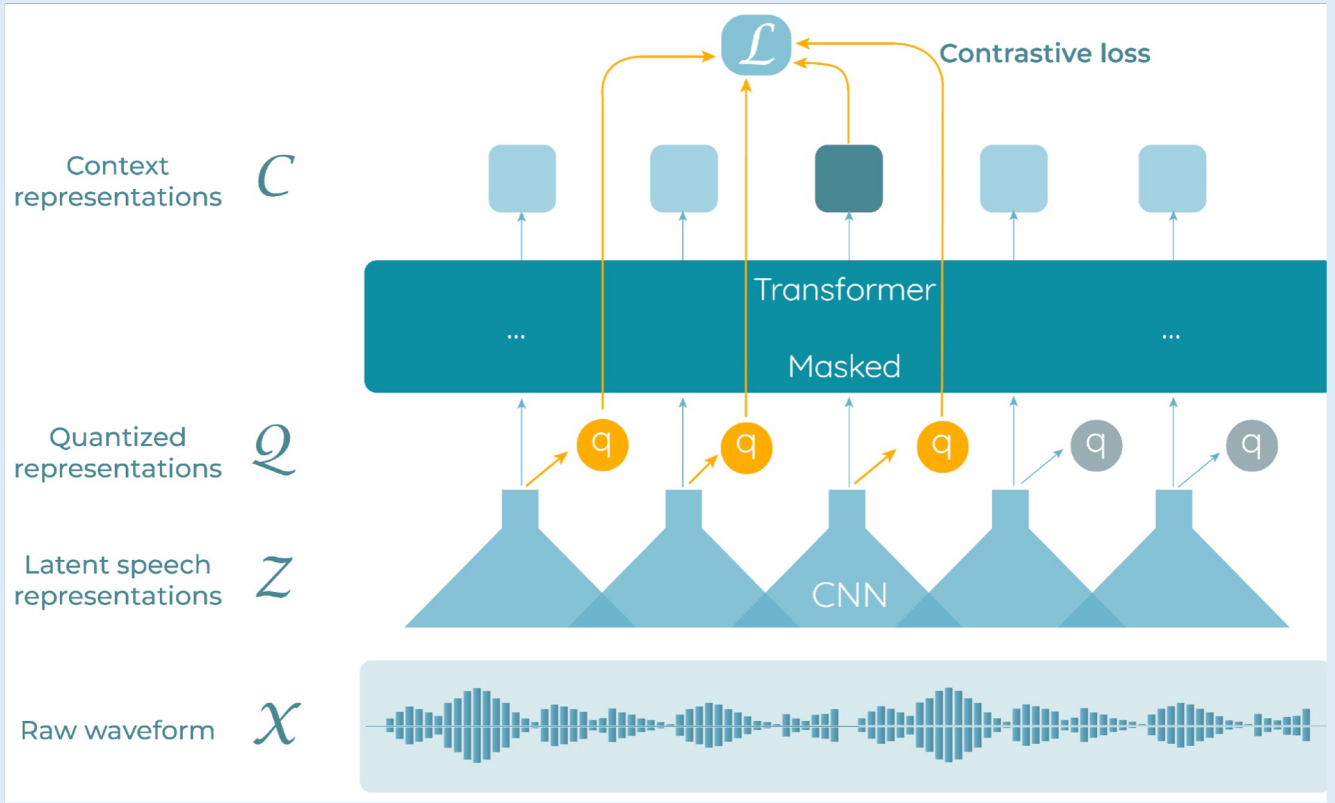
- The onset of a gesture phrase precedes the onset of speech
  - 200 –500 ms
  - There is no perfect alignment between these two communication cues
  - Familiarity with the lexical concepts and common ground plays a role
- Beat gestures co-occur (in a close synchrony) with stressed syllables
- Speech and gesture coordination is not only an intra-speaker phenomenon but could be inter-speaker: gestural alignment





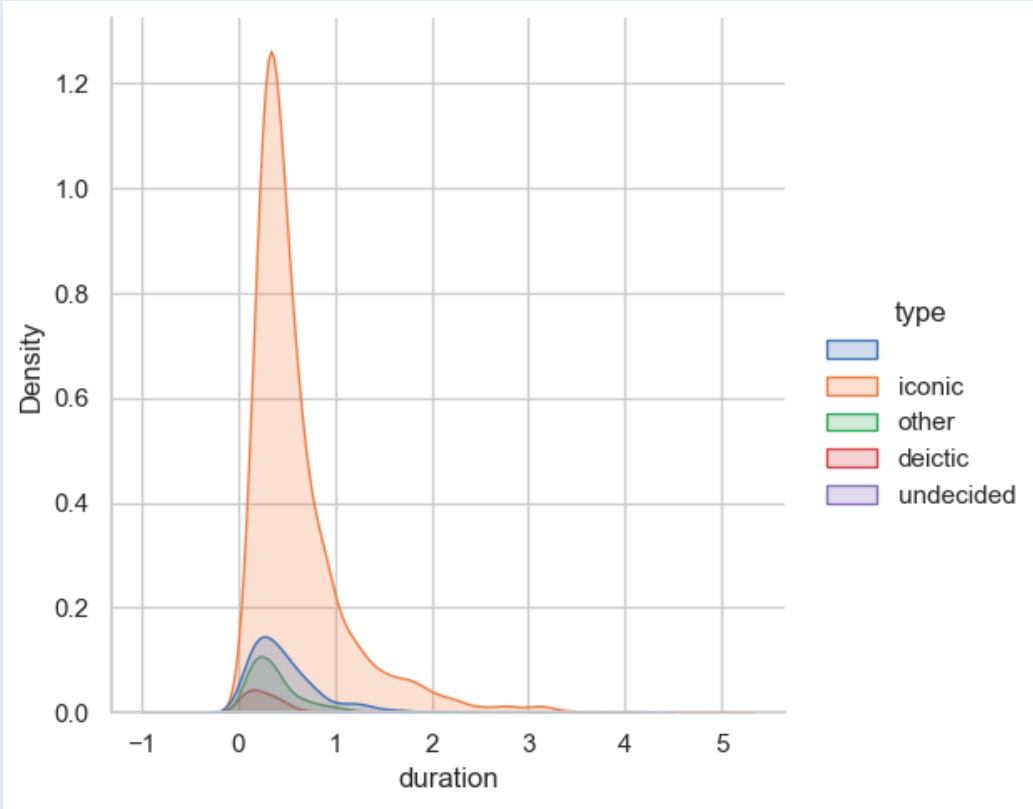
Speech Embeddings: WAV2VEC2

- Pretrained model using similar objective used in language modeling, e.g., the one in BERT
  - **Wav2Vec2-XLSR-300:** cross-lingual speech representations, pre-trained from the raw waveform of speech in multiple languages



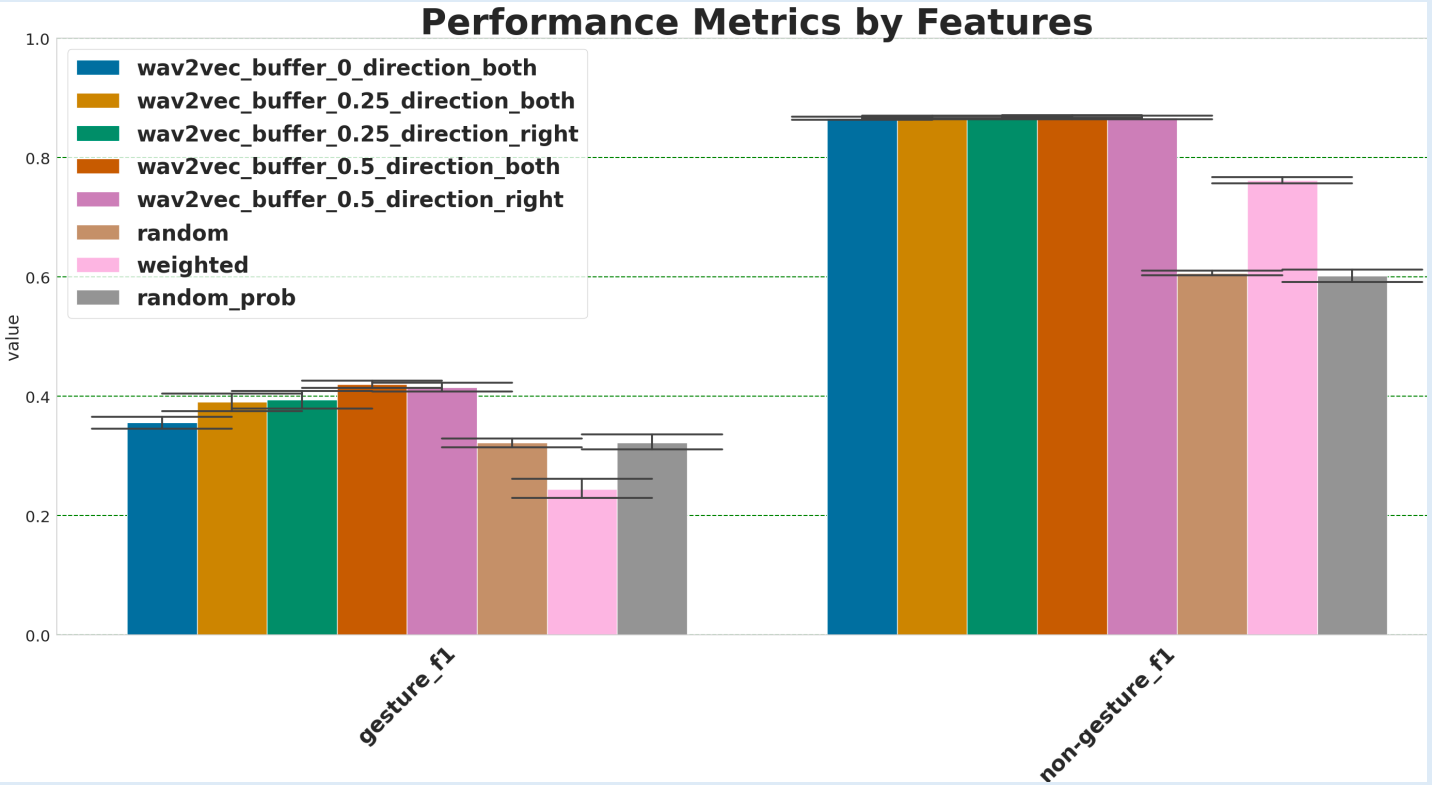
Can We Detect Gestures Using Speech?

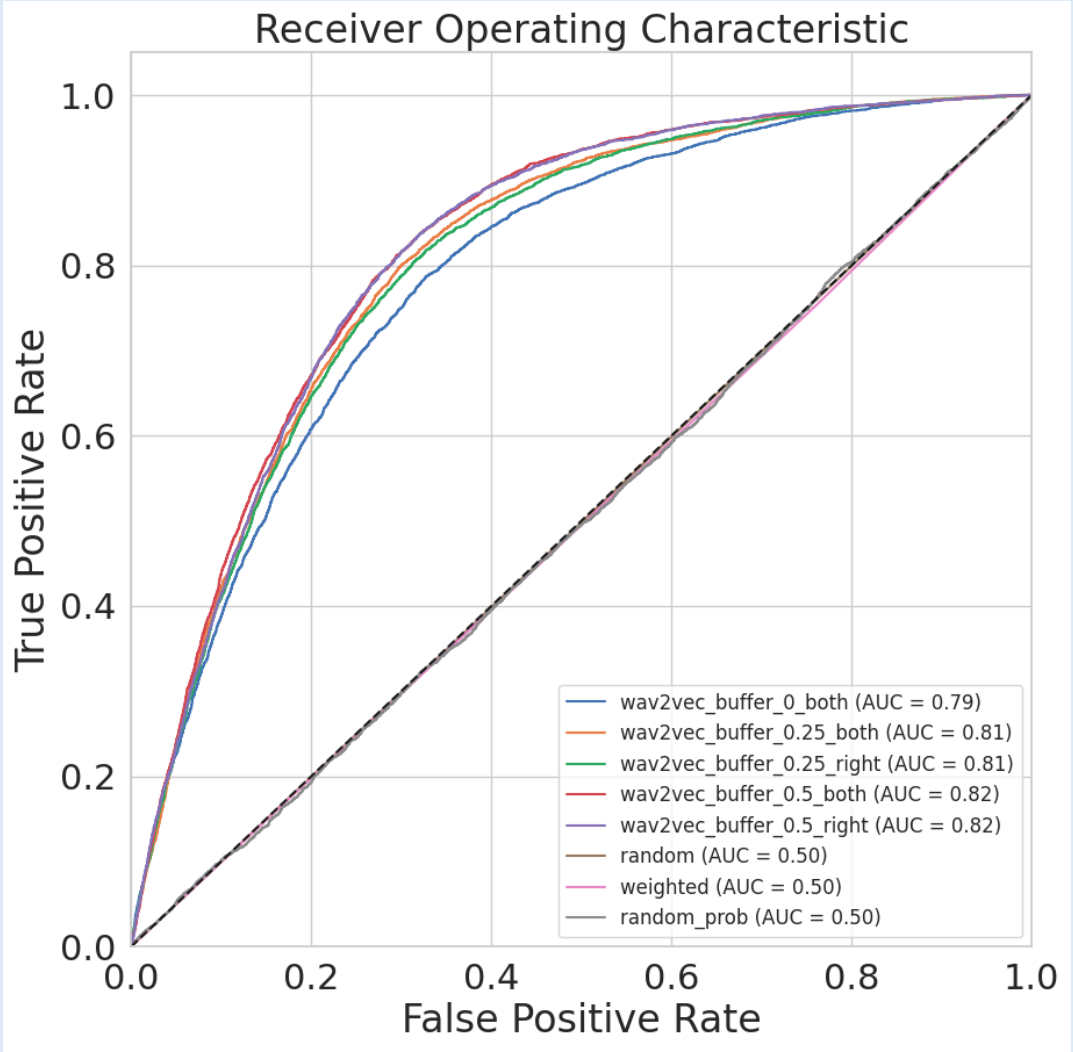
- We use a clean dataset
- → No sliding windows & no sequences
- Samples for gestures:
  - Annotated strokes of gestures: > 5K segmented time windows
- Samples of other movements (non-gestures)
  - Randomly segmented time windows: 15K
  - Excluding any time window that overlaps with a stroke of a gesture
- Divide data into ten speaker-disjoint folds for cross-validation
- Train a binary classifier: gesture or not
  - Logistic regression



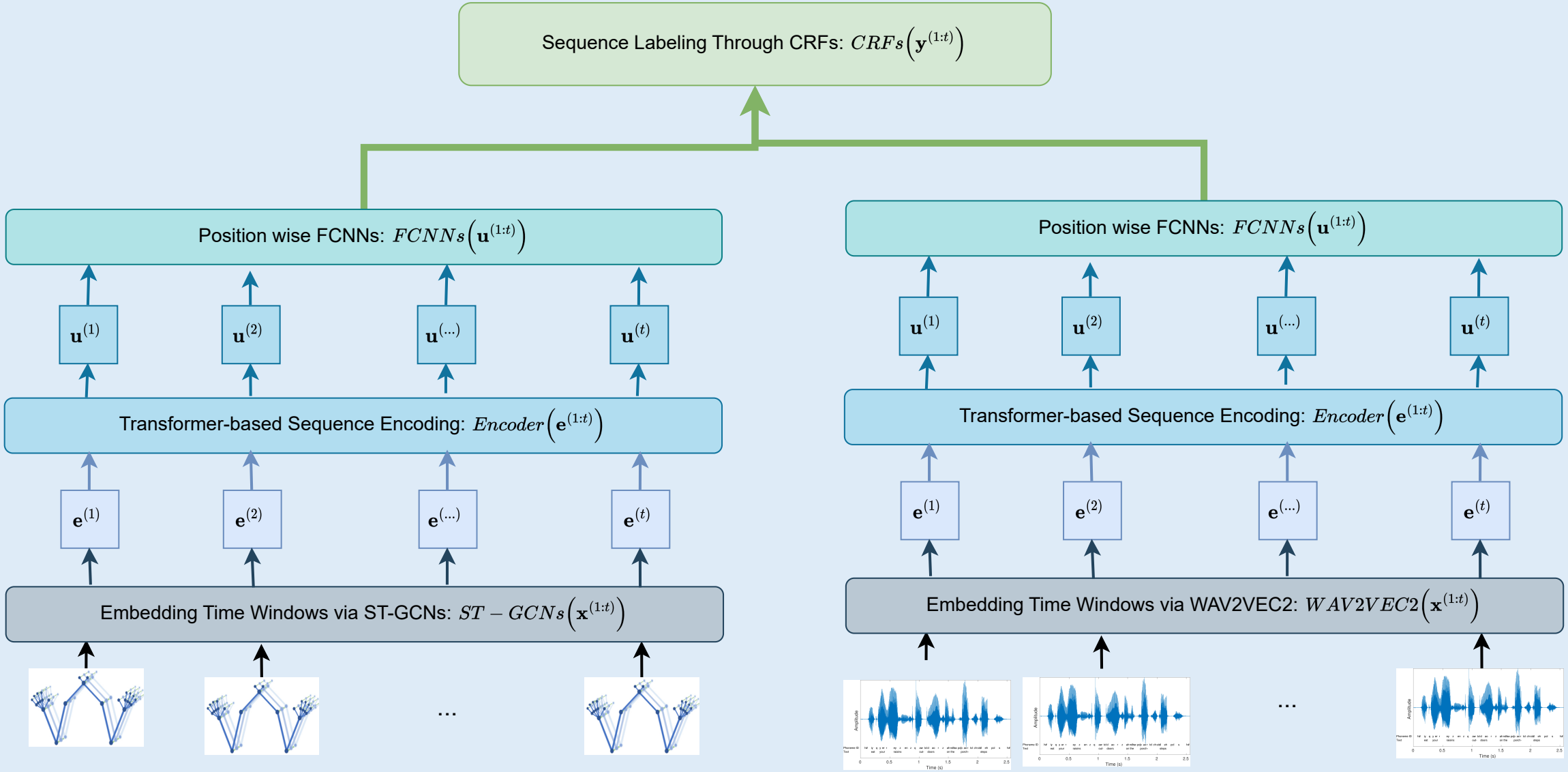
Results: F1-Score & ROC Curve → Speech Only

- WAV2VEC2 features are pooled for each segmented time window
  - With or without a buffer (centered or to the right alone)



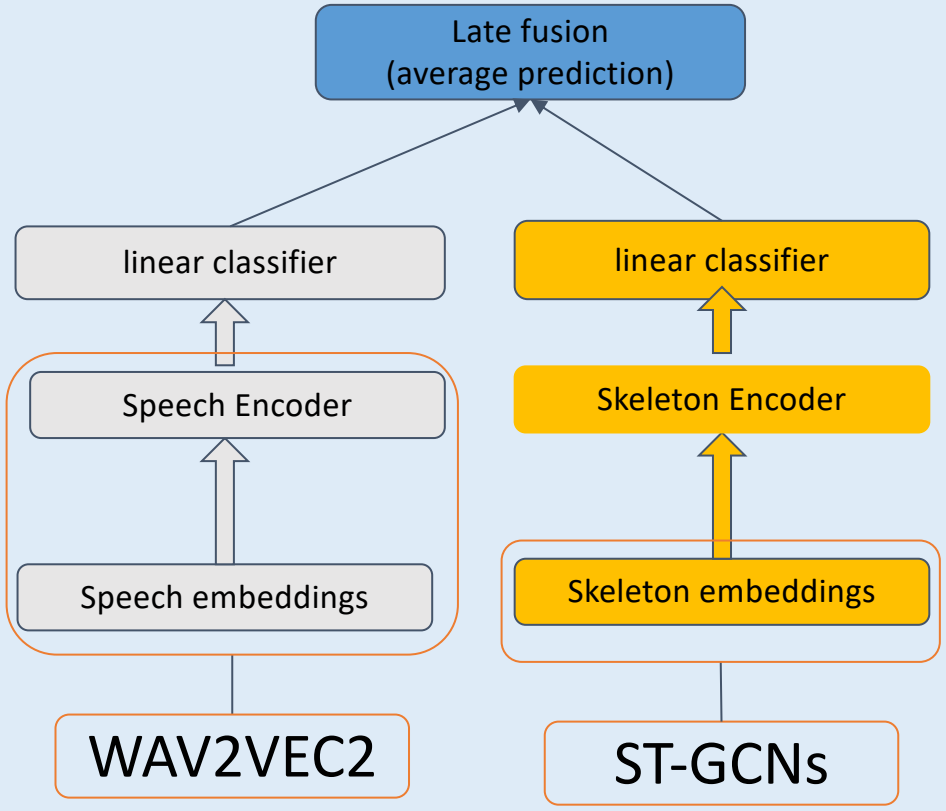


Late Fusion: Combining Speech & Skeletal Models



Late Fusion: Sliding Window Approach

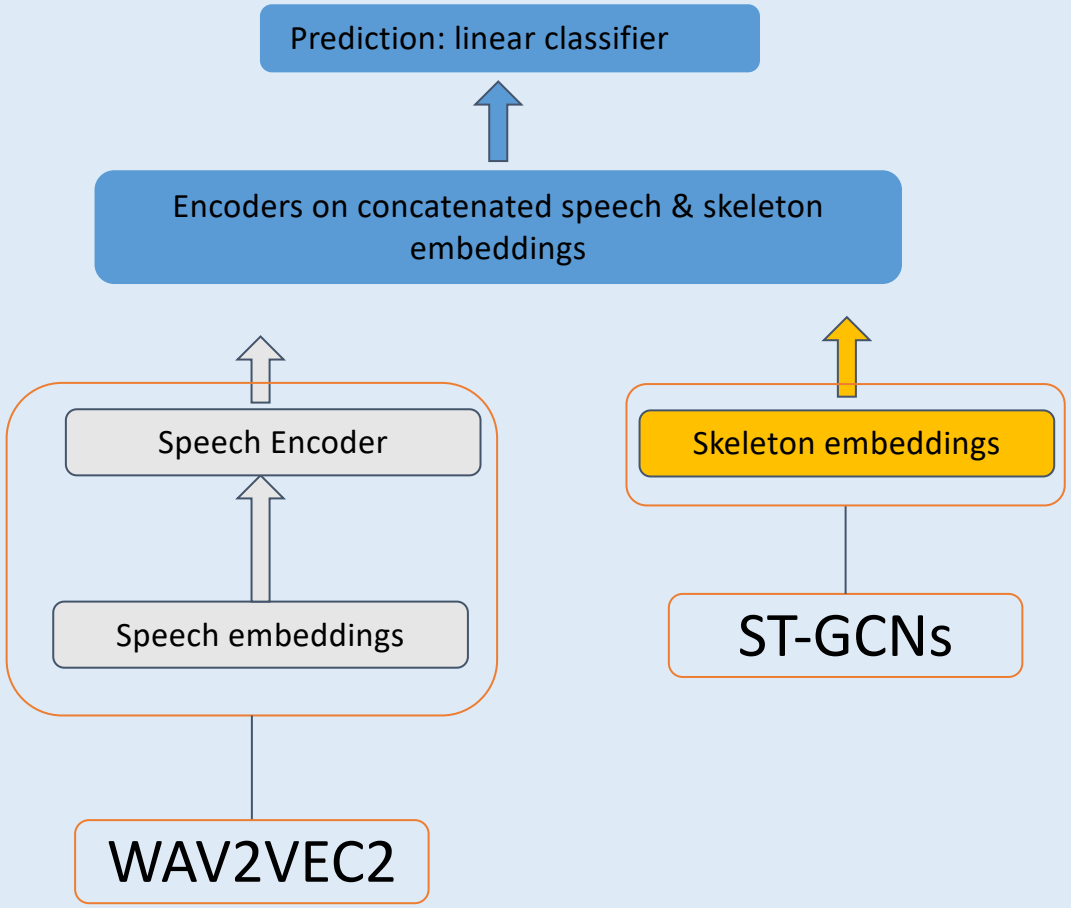
- Separate embedding models for speech and visual cues
  - Separate encoders and classifiers
- Binary classification (neutral and stroke)
  - Stroke → overlapping with stroke (i.e., >50%)
  - Skeleton + WAV2VEC2: **66.7.0%**
  - Skeleton only: **66.2%**
- In this way, the speech model **improves** the detection performance, but not significantly!



Early Fusion

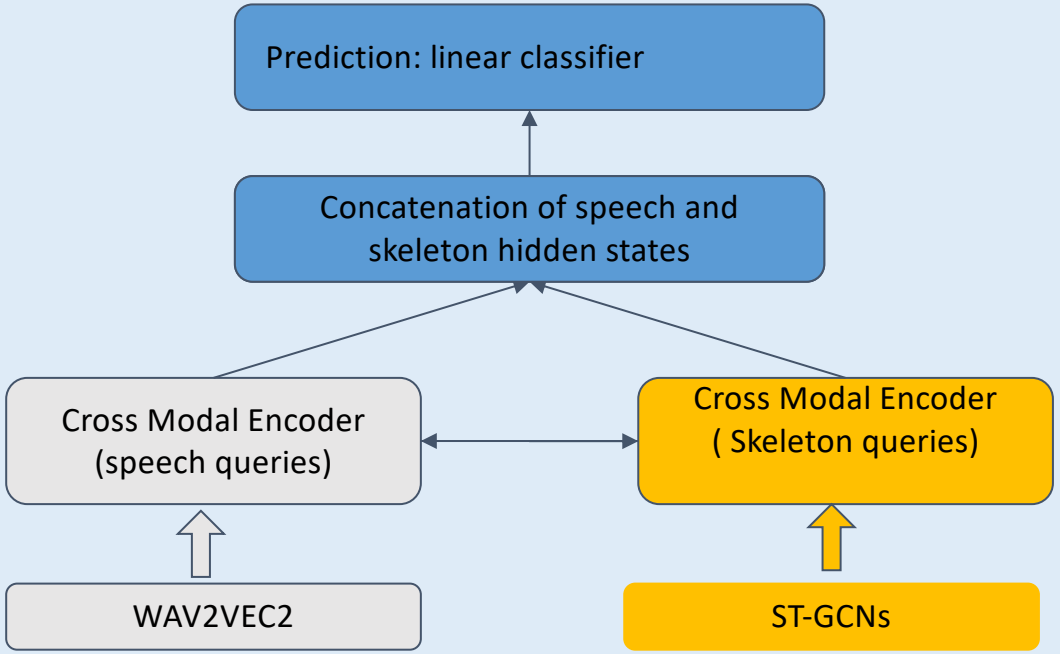
- Separate embedding models for speech and visual cues

- Transformer encoder on the concatenated speech and skeletal embeddings



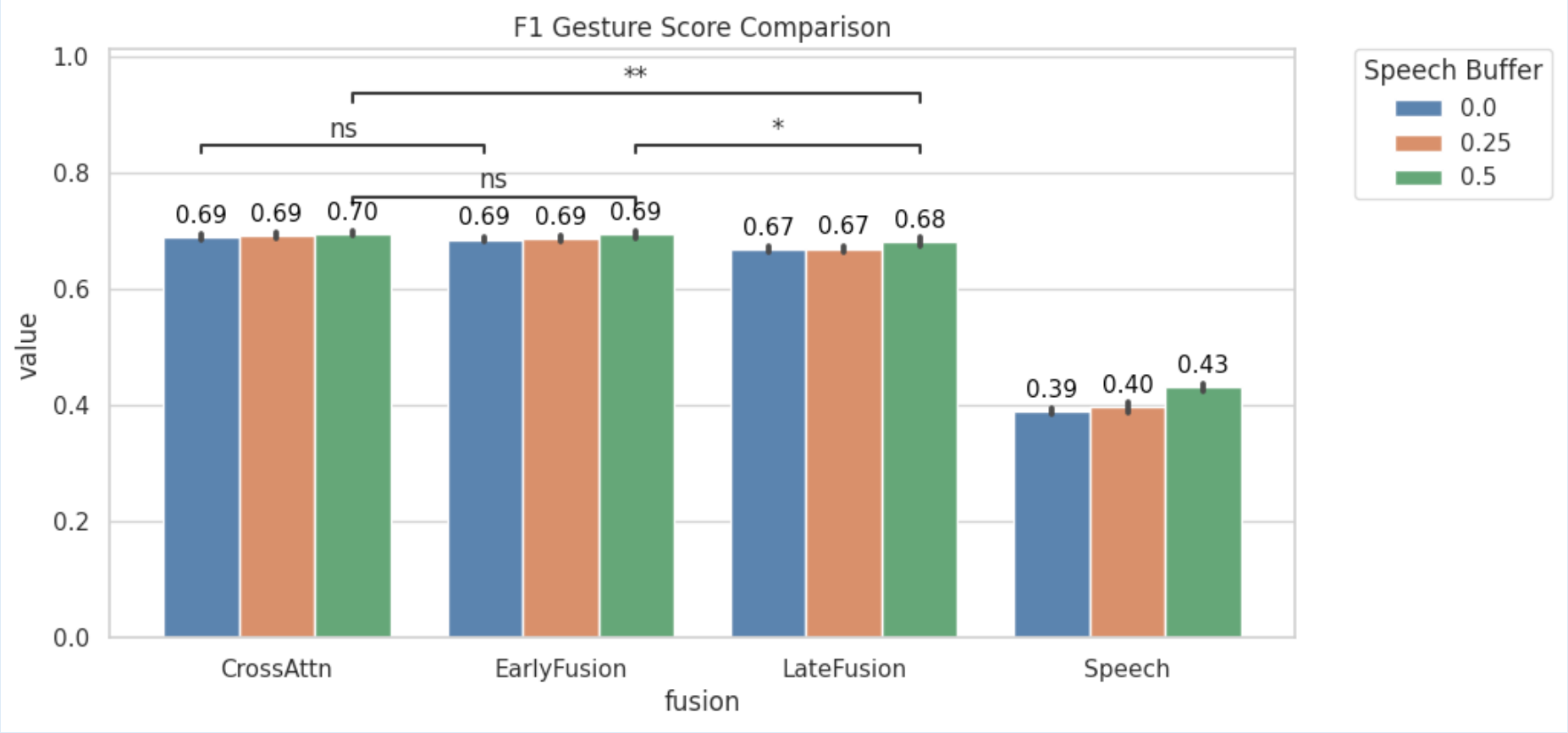
Architecture Based on Cross-modal Attention Inspired by LXMERT (Tan & Bansal, 2019)

- Multihead cross-modal attention
  - Queries from one modality
    - Keys and values from the other modality
- Linear classifier on the concatenated embeddings



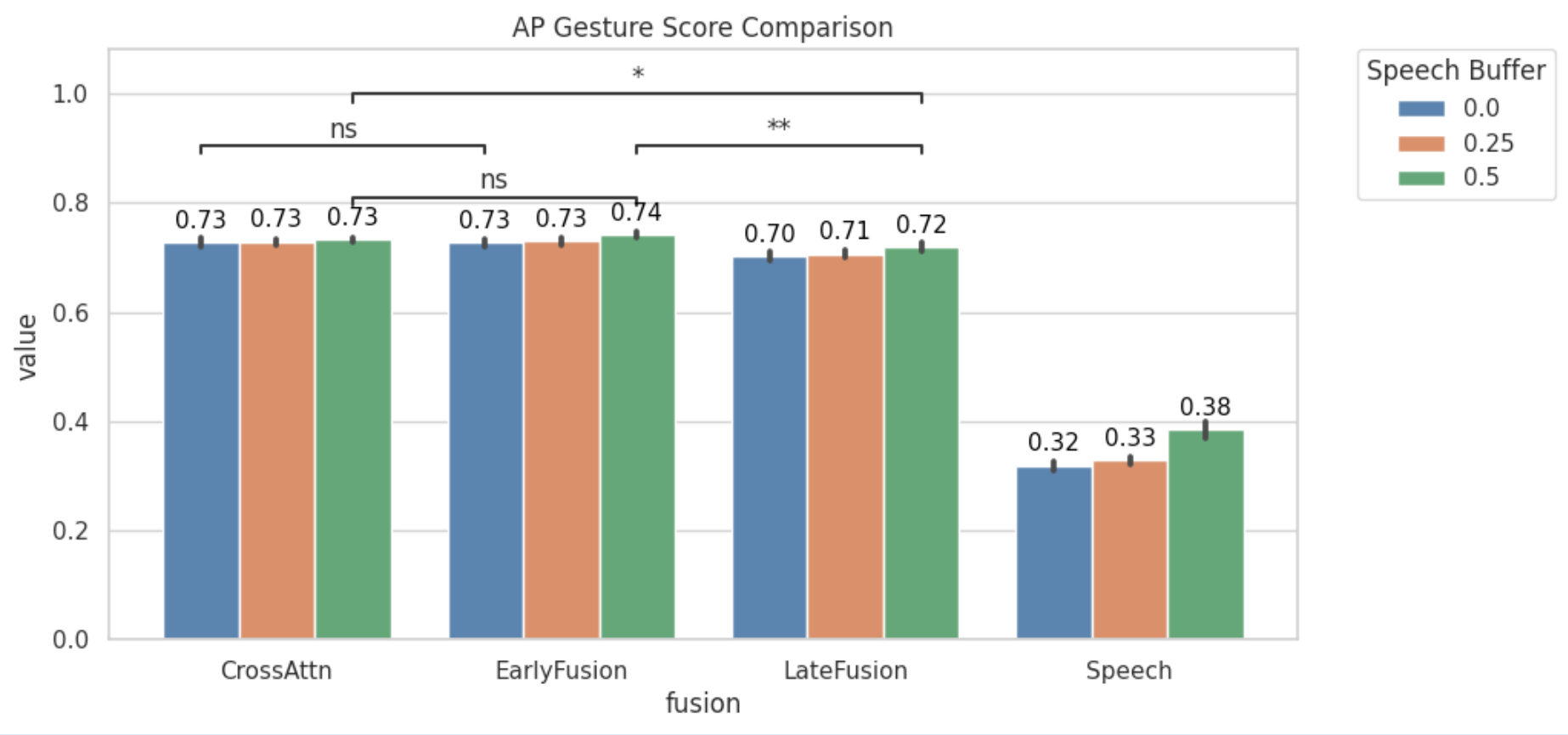
Results: Gesture Stroke Detection - F1

- Multimodal fusion based on cross-attention models outperforms early & late fusion



Results: Gesture Stroke Detection - Average Precision

- Speech buffer matters less in the cross-attention model & early Fusion Models



Conclusions

- We Proposed a framework to align speech and gestures.
- Multimodal integration and alignment through early and cross-attention models give better performance than those that do not integrate both modalities.
- Speech buffer helps unimodal speech and late fusion models

The Dialogue Modelling Group

https://dmg-illc.github.io/dmg/

Dialogue Modelling Group

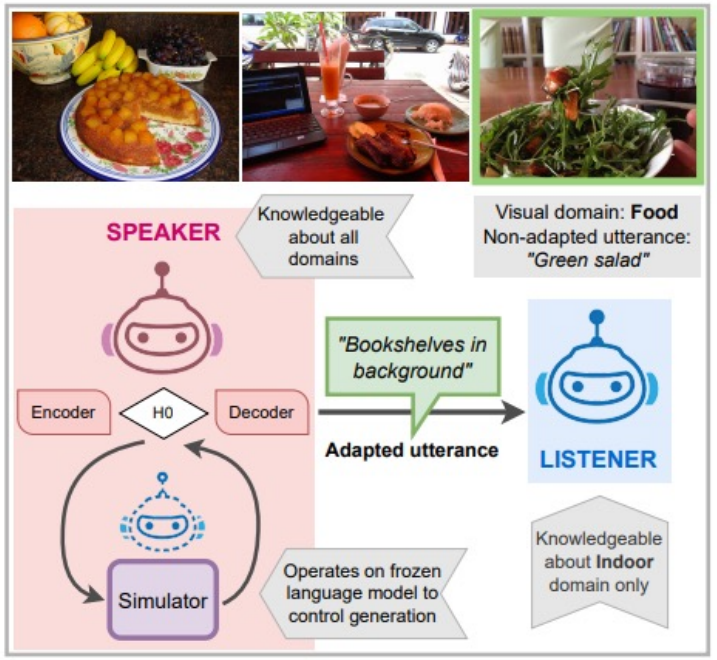
- Dialogue partner's knowledge aware modelling

Speaking the Language of Your Listener: Audience-Aware Adaptation via Plug-and-Play Theory of Mind

Ece Takmaz[◁*], Nicolo' Brandizzi[◇*], Mario Giulianelli[◁], Sandro Pezzelle[◁], Raquel Fernánd
[◁]University of Amsterdam [◇]Sapienza University of Rome
{ece.takmaz|m.giulianelli|s.pezzelle|raquel.fernandez}@uva.nl
brandizzi@diag.uniroma1.it

Abstract

Dialogue participants may have varying levels of knowledge about the topic under discussion. In such cases, it is essential for speakers to adapt their utterances by taking their audience into account. Yet, it is an open question how such adaptation can be modelled in computational agents. In this paper, we model a visually grounded referential game between a knowledgeable speaker and a listener with more limited visual and linguistic experience. Inspired by psycholinguistic theories, we endow our speaker with the ability to adapt its referring expressions via a simulation module that monitors the effectiveness of planned utterances



Dialogue Modelling Group

- Evaluations metrics

GROOViST: A Metric for Grounding Objects in Visual Storytelling

Aditya K Surikuchi
University of Amsterdam
a.k.surikuchi@uva.nl

Sandro Pezzelle, Raquel Fernández
ILLC, University of Amsterdam
{s.pezzelle, raquel.fernandez}@uva.nl

Abstract

A proper evaluation of stories generated for a sequence of images—the task commonly referred to as visual storytelling—must consider multiple aspects, such as coherence, grammatical correctness, and visual grounding. In this work, we focus on evaluating the degree of grounding, that is, the extent to which a story is about the entities shown in the images. We analyze current metrics, both designed for this purpose and for general vision-text alignment. Given their observed shortcomings, we propose a novel evaluation tool, GROOViST, that accounts for cross-modal dependencies, *temporal misalignments* (the fact that the order in which entities appear in the story and the image se-

1) there was lots to see and do at the festival , including listening to unusual instruments . 2) many stalls had handmade clothing and one even had dresses specifically for little girls . 3) as part of the festival grounds , there were also numerous sculptures that one could touch . 4) many stalls were adorned with handmade glass bottles . 5) by midday thousands were in attendance , the biggest turn out yet !

Figure 1: One story and corresponding image sequence from the VIST dataset. Noun phrases in green contribute positively to the grounding score by GROOViST; those in red contribute negatively. The GROOViST score for this sample is 0.855, i.e., our metric considers it as well-grounded (within range: $[-1, 1]$). Best viewed in color.

appropriate—they indeed poorly correlate with

Thank you for your attention!

- Questions?

Readings

1. Ter Bekke, M., Drijvers, L., & Holler, J. (2020). The predictive potential of hand gestures during conversation: An investigation of the timing of gestures in relation to speech.
2. Donnellan, E., Özder, L. E., Man, H., Grzyb, B., Gu, Y., & Vigliocco, G. (2022, July). Timing relationships between representational gestures and speech: A corpus-based investigation. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 44, pp. 2052-2058). University of California.
3. Wagner, P., Malisz, Z., & Kopp, S. (2014). Guest Editorial: Gesture and speech in interaction: An overview. *Speech Communication*, *57*, 209-232.
4. *Rajan, V., Brutti, A., & Cavallaro, A. (2022, May). Is cross-attention preferable to self-attention for multi-modal emotion recognition?. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 4693-4697). IEEE.*
5. *Holler, J., & Wilkin, K. (2011). Co-speech gesture mimicry in the process of collaborative referring during face-to-face dialogue. Journal of Nonverbal Behavior, 35, 133-153.*
6. *Holler, J., & Levinson, S. C. (2019). Multimodal language processing in human communication. Trends in Cognitive Sciences, 23(8), 639-652.*
7. *Lücking, Andy, et al. "The Bielefeld speech and gesture alignment corpus (SaGA)." LREC 2010 workshop: Multimodal corpora–advances in capturing, coding and analyzing multimodality. 2010.*
8. *an, Sijie, Yuanjun Xiong, and Dahua Lin. "Spatial temporal graph convolutional networks for skeleton-based action recognition." Thirty-second AAAI conference on artificial intelligence. 2018.*
9. *Ghaleb, Esam, et al. "Skeleton-based explainable bodily expressed emotion recognition through graph convolutional networks." 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021). IEEE, 2021.*
10. *Jiang, Songyao, et al. "Skeleton aware multi-modal sign language recognition." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.*
11. *Sun, Ke, et al. "Deep high-resolution representation learning for human pose estimation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019*
12. *Ozge Mercanoglu Sincan and Hacer Yalim Keles. AUTSL: A large scale multi-modal turkish sign language dataset and baseline methods. IEEE Access, 8:181340–181355, 2020. 6*
13. *Köpüklü, Okan, et al. "Online dynamic hand gesture recognition including efficiency analysis." IEEE Transactions on Biometrics, Behavior, and Identity Science 2.2 (2020): 85-97.*