Observing Dialogue in Therapy: Categorizing and Forecasting Behavioral Codes

Jie Cao, Michael Tanana, Zac E. Imel, Eric Poitras, David C. Atkins, Vivek Srikumar

*Tanana, Imel are co-founders and minority equity stakeholders of a technology company – Lyssn.io that is focused on developing computational models that quantify aspects of patient-provider interactions in psychotherapy.

Can We Obtain Expertise in Mental Health Treatment?

Expertise is Developed When:

”The environment is predictable with explicit outcomes”

“There is an opportunity to learn based on quality information”

(Tracey, Wampold, Lichtenberg & Goodyear (2014) summarizing Kahneman and Klein (2009))

This paper

1. Motivation for real-time feedback in therapy
2. Defines two tasks: categorizing and forecasting MISC codes
3. Systematically tests modeling choices
4. Proposes neural models that outperform several baselines

What is Motivational Interviewing?

Evidence-based form of psychotherapy

Understanding client perspective to motivate change

Utterance level Behavioral Codes

Code	Count	Description	Examples
Client Behavioral Codes			
FN	47715	Follow/ Neutral: unrelated to changing or sustaining behavior.	"You know, I didn't smoke for a while." "I have smoked for forty years now."
CT	5099	Utterances about changing unhealthy behavior.	"I want to stop smoking."
ST	4378	Utterances about sustaining unhealthy behavior.	"I really don't think I smoke too much."

Utterance level Behavioral Codes

Code	Count	Description	Examples
Client Behavioral Codes			
FN	47715	Follow/ Neutral: unrelated to changing or sustaining behavior.	"You know, I didn't smoke for a while." "I have smoked for forty years now."
CT	5099	Utterances about changing unhealthy behavior.	"I want to stop smoking."
ST	4378	Utterances about sustaining unhealthy behavior.	"I really don't think I smoke too much."
Therapist Behavioral Codes			
FA	17468	Facilitate conversation	"Mm Hmm.", "OK.","Tell me more."
GI	15271	Give information or feedback.	"I'm Steve.", "Yes, alcohol is a depressant."
RES	6246	Simple reflection about the clients most recent utterance.	C: "I didn't smoke last week" T: "Cool, you avoided smoking last week."
REC	4651	Complex reflection based on a client's history or the broader conversation.	C: "I didn't smoke last week." T: "You mean things begin to change".
QUC	5218	Closed question	"Did you smoke this week?"
QUO	4509	Open question	"Tell me more about your week."
MIA	3869	Other MI adherent,*e.g.*, affirmation, advising with permission, etc.	"You've accomplished a difficult task." "Is it OK if I suggested something?"
MIN	1019	MI non-adherent, *e.g.*, confrontation, advising without permission, etc.	"You hurt the baby's health for cigarettes?" "You ask them not to drink at your house."

Why real-time feedback?

1. Post-hoc analysis does not always help
   a. Feedback is not in real-time, cannot correct errors from hours ago
   b. Less helpful for therapist training

2. Real-time feedback can...
   a. monitor fidelity to therapy standards
   b. alert the therapist to potentially important cues from the client
   c. offer suggestions to trainees

Two Tasks

1. **Categorization**: Monitoring an ongoing session by predicting MISC labels for therapist and client utterances as they are made.

2. **Prediction:** Given a dialogue history, forecasting the MISC label for the next utterance, thereby both alerting or guiding therapists

An example session

Therapist: Have you used any drugs recently?	Closed question
Client: I had stopped, but recently relapsed…	Follow Neutral
Therapist: You'll suffer if you keep this up.	MI Non-adherent
Client: Sorry, I just want to quit.	Change Talk

Data

353 psychotherapy sessions

Annotated at the utterance level with MISC codes

243 training sessions/ 110 testing

Splits used in Can et al. (2015); Tanana et al. (2016)

24 of the training sessions formed the dev set

Modeling dialogue observers

Modeling dialogue observers

Given a history of utterances, we need to predict the MISC label for:

- The last one (Categorization)
- The next one (Forecasting)

We have four modeling questions to address:

1. Encode words and utterances	2. Discover discriminative words	3. Use (only) relevant utterances	4. Address label imbalance
Hierarchical GRU	Word level attention	Utterance level attention	Focal loss

Modeling dialogue observers

Given a history of utterances, we need to predict the MISC label for:

- The last one (Categorization)
- The next one (Forecasting)

We have four modeling questions to address:

1. Encode words and utterances	2. Discover discriminative words	3. Use (only) relevant utterances	4. Address label imbalance
Hierarchical GRU	Word level attention	Utterance level attention	Focal loss

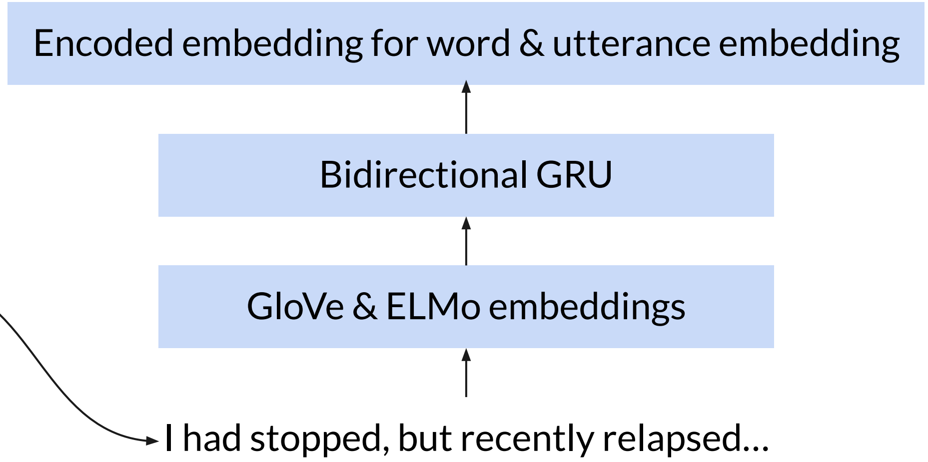
Encoding words & utterances: Hierarchical GRU

Therapist: Have you used any drugs recently?

Client: I had stopped, but recently relapsed…

Therapist: You’ll suffer if you keep this up.

Client: Sorry, I just want to quit.



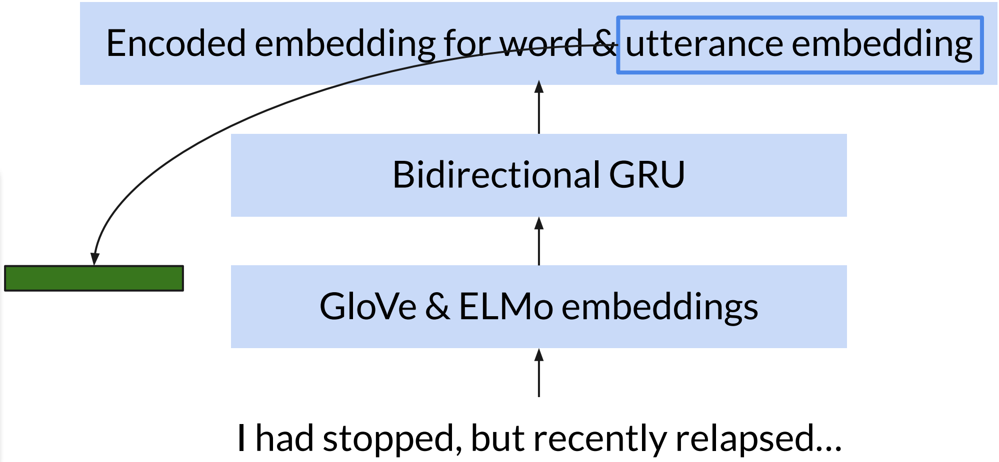
Encoding words & utterances: Hierarchical GRU

Therapist: Have you used any drugs recently?

Client: I had stopped, but recently relapsed…

Therapist: You'll suffer if you keep this up.

Client: Sorry, I just want to quit.



Encoding words & utterances: Hierarchical GRU

Therapist: Have you used any drugs recently?

Client: I had stopped, but recently relapsed…

Therapist: You’ll suffer if you keep this up.

Client: Sorry, I just want to quit.

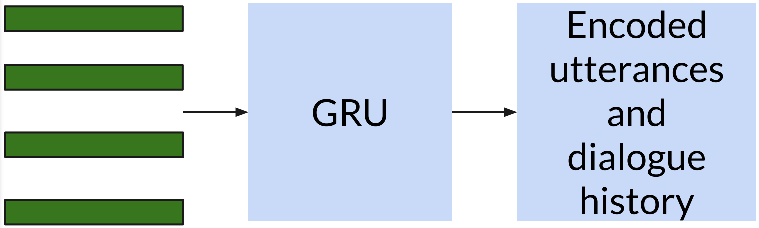
Encoding words & utterances: Hierarchical GRU

Therapist: Have you used any drugs recently?

Client: I had stopped, but recently relapsed…

Therapist: You'll suffer if you keep this up.

Client: Sorry, I just want to quit.



This forms the general scaffolding for <u>*all*</u> our models.

Modeling dialogue observers

Given a history of utterances, we need to predict the MISC label for:

- The last one (Categorization)
- The next one (Forecasting)

We have four modeling questions to address:

1. Encode words and utterances	2. Discover discriminative words	3. Use (only) relevant utterances	4. Address label imbalance
Hierarchical GRU	Word level attention	Utterance level attention	Focal loss

Do we really need hierarchical attention for our tasks?

Attending to words and utterances

- Attention mechanisms built over the encoded word and utterance vectors
- Validation set to find best attention mechanism, if necessary
    - (We will see in results that they are not always necessary)

2. Discover discriminative words

3. Use (only) relevant utterances

Word level attention

Gated Match GRU
Based on Match LSTM (Wang et al 2017)

Utterance level attention

Multi-headed attention, with 4 heads, 2 hops
Using transformers (Vaswani et al 2017)

See paper for details

Modeling dialogue observers

Given a history of utterances, we need to predict the MISC label for:

- The last one (Categorization)
- The next one (Forecasting)

We have four modeling questions to address:

1. Encode words and utterances

2. Discover discriminative words

3. Use (only) relevant utterances

4. Address label imbalance

Hierarchical GRU

Word level attention

Utterance level attention

Focal loss

Addressing label imbalance with focal loss

- Problem: Some labels (e.g. Change Talk, Sustain Talk, MI Non-adherent) are crucial, but rare in the data
    - Standard loss will be dominated by large number of easy labels

- Focal loss extends standard cross-entropy:

(Lin et al 2017)

$$\mathrm{FL}(p_t) = -\alpha_t (1 - p_t)^{\gamma} \log(p_t)$$

Addressing label imbalance with focal loss

- Problem: Some labels (e.g. Change Talk, Sustain Talk, MI Non-adherent) are crucial, but rare in the data
    - Standard loss will be dominated by large number of easy labels

- Focal loss extends standard cross-entropy:

$$\mathrm{FL}(p_t) = -\boxed{\alpha_t}(1 - p_t)^\gamma \log(p_t)$$

A label specific scaling factor that can down-weight less important labels

Addressing label imbalance with focal loss

- Problem: Some labels (e.g. Change Talk, Sustain Talk, MI Non-adherent) are crucial, but rare in the data
  - Standard loss will be dominated by large number of easy labels

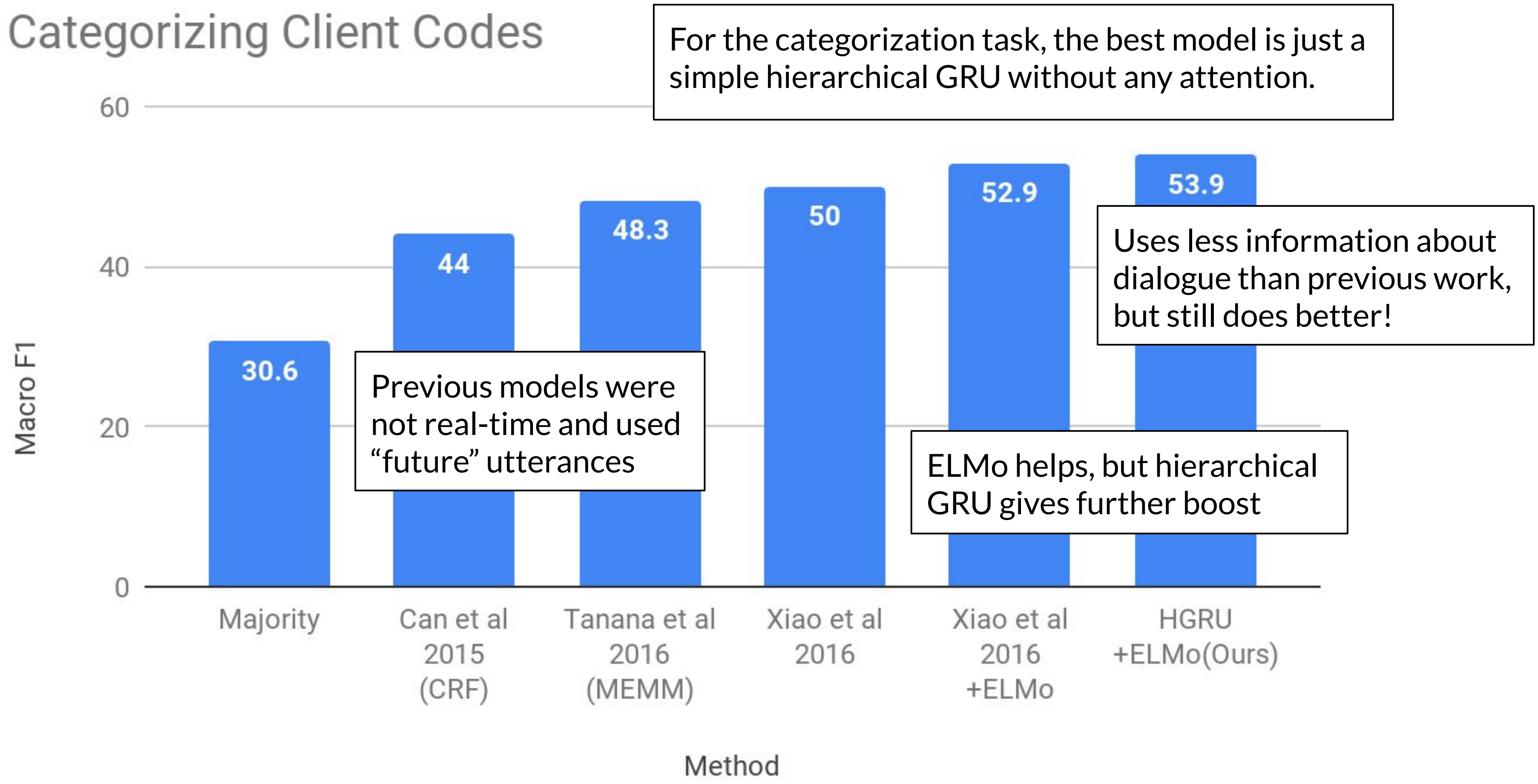
- Focal loss extends standard cross-entropy:

$$\mathrm{FL}(p_t) = -\alpha_t \boxed{(1 - p_t)^\gamma} \log(p_t)$$

A label specific scaling factor that can down-weight less important labels

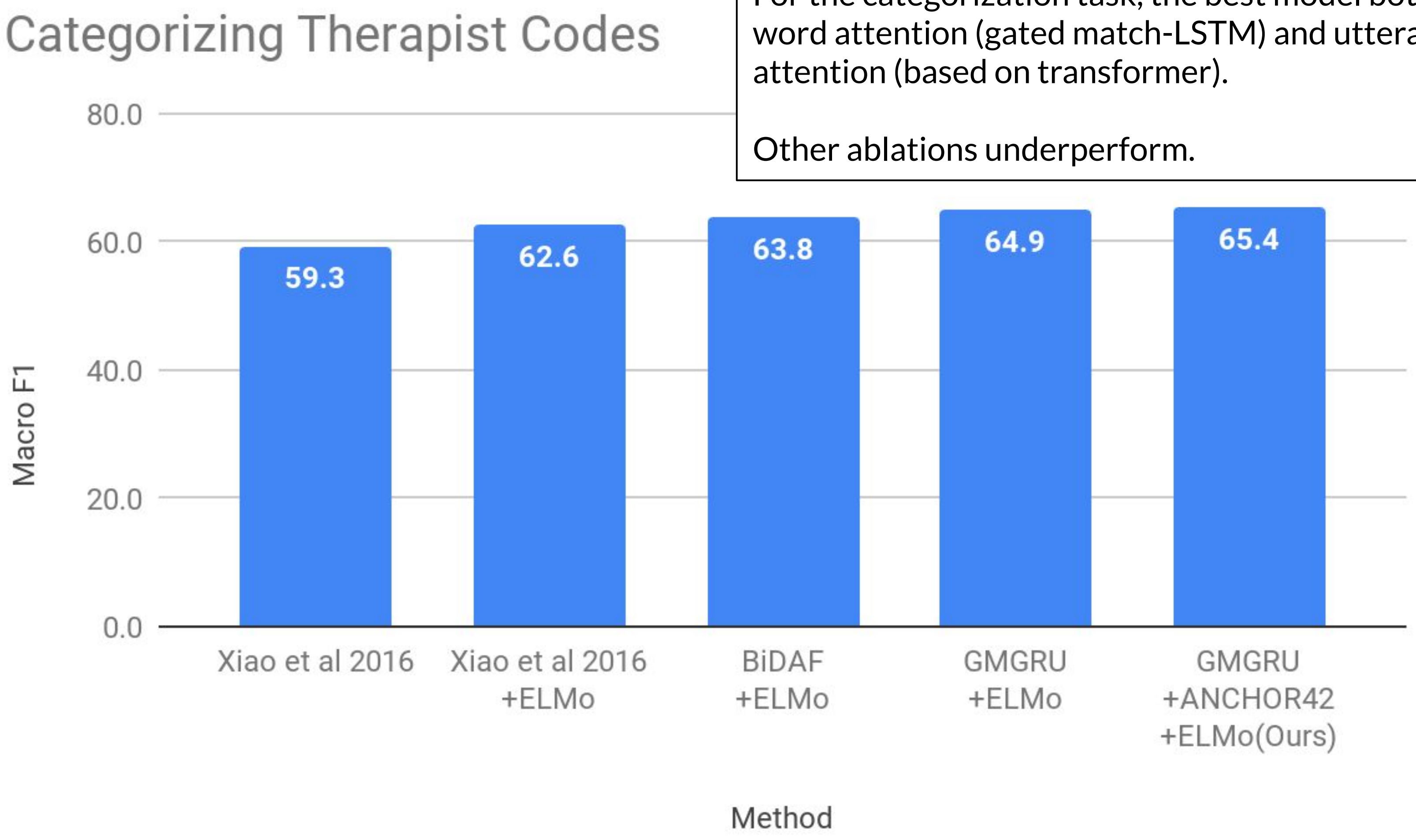
A multiplier that ensures that easy-to-predict labels have low loss

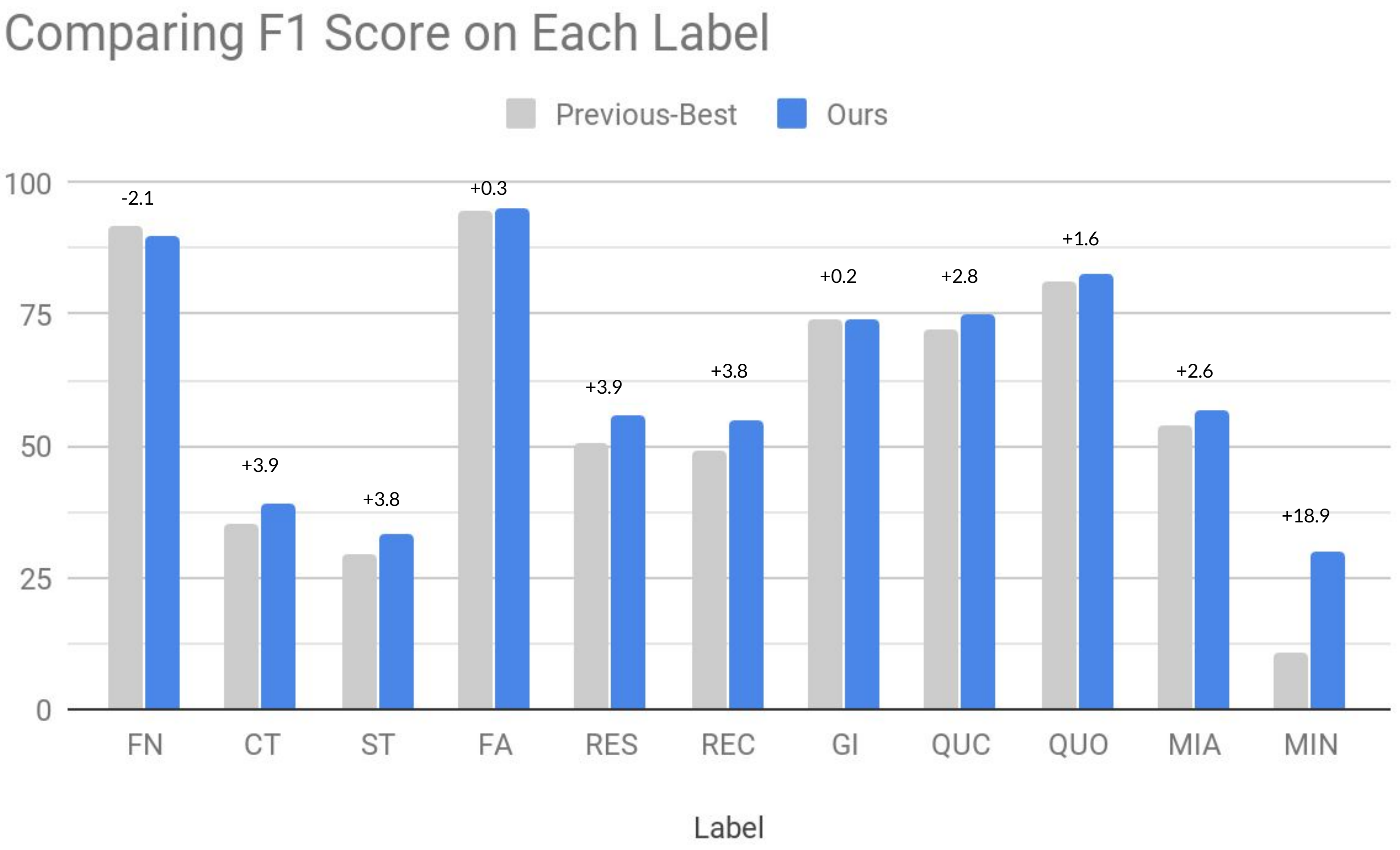
Results: Categorization Task


Categorizing Client Codes
For the categorization task, the best model is just a simple hierarchical GRU without any attention.
60
40
20
0
Macro F1
30.6
44
48.3
50
52.9
53.9
Previous models were not real-time and used "future" utterances
Uses less information about dialogue than previous work, but still does better!
ELMo helps, but hierarchical GRU gives further boost
Majority
Can et al 2015 (CRF)
Tanana et al 2016 (MEMM)
Xiao et al 2016
Xiao et al 2016 +ELMo
HGRU +ELMo(Ours)
Method

For the categorization task, the best model both word attention (gated match-LSTM) and utterance attention (based on transformer).

Other ablations underperform.





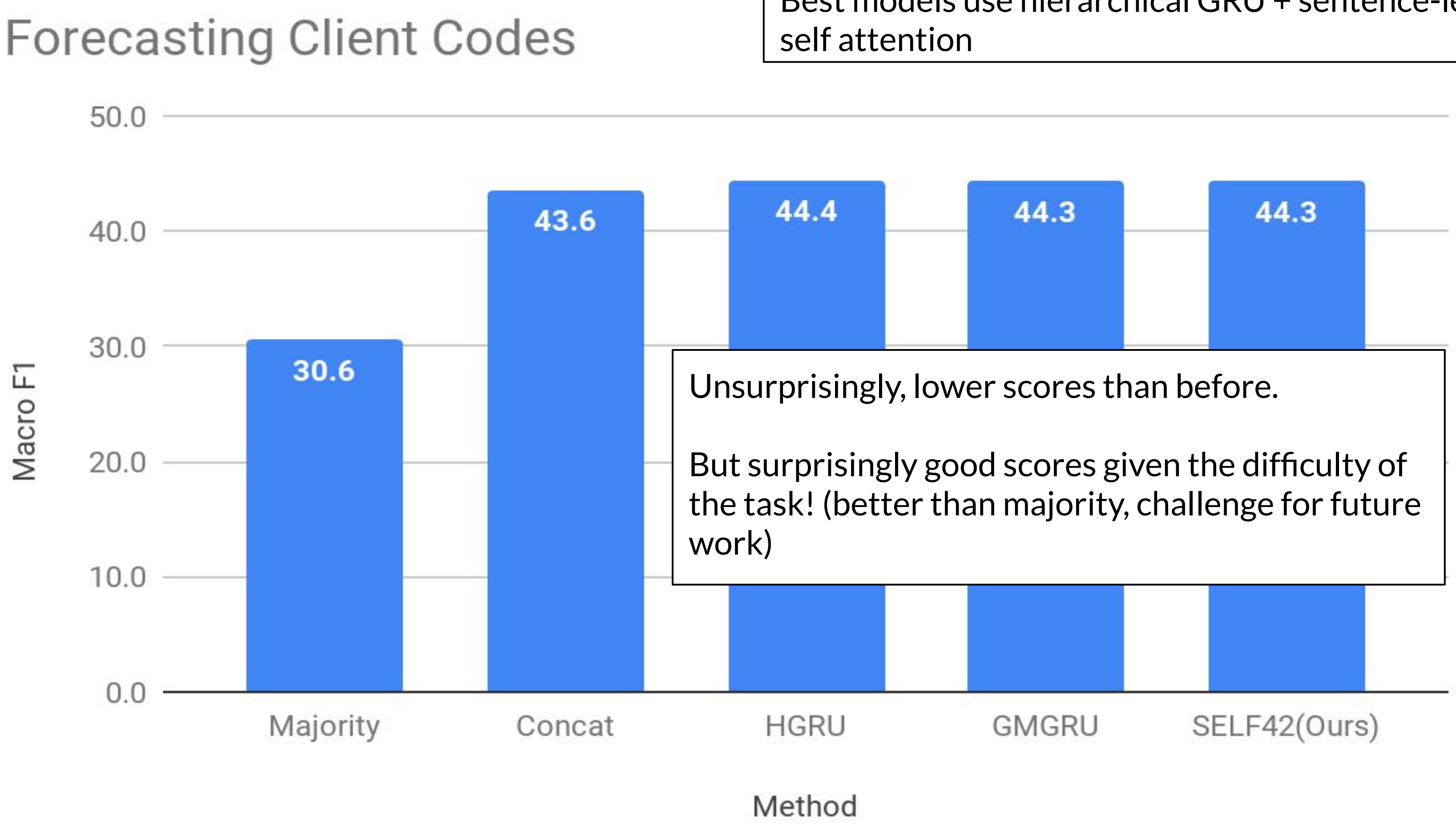
Comparing F1 Score on Each Label
Previous-Best
Ours
100
75
50
25
0
-2.1
+3.9
+3.8
+0.3
+3.9
+3.8
+0.2
+2.8
+1.6
+2.6
+18.9
FN
CT
ST
FA
RES
REC
GI
QUC
QUO
MIA
MIN
Label

Results: Forecasting Task

Recall that this task calls for predicting a label *before* seeing the utterance for which the label applies!

No previous baselines. So we will see comparisons to ablations.

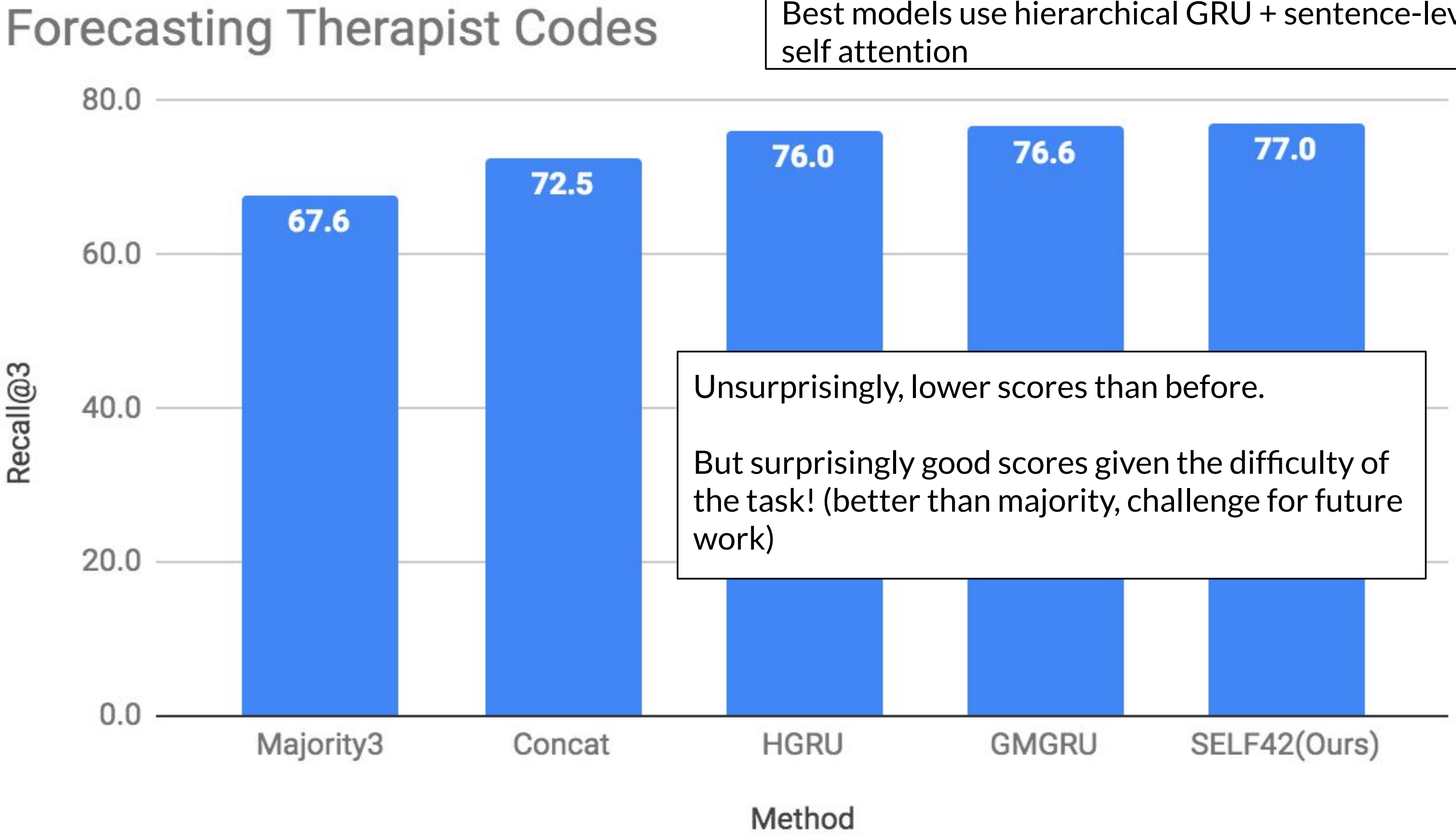
Best models use hierarchical GRU + sentence-level self attention



Unsurprisingly, lower scores than before.

But surprisingly good scores given the difficulty of the task! (better than majority, challenge for future work)

Best models use hierarchical GRU + sentence-level self attention



Unsurprisingly, lower scores than before.

But surprisingly good scores given the difficulty of the task! (better than majority, challenge for future work)

What else have we learned: Analysis

1. Dialogue context helps to some extent
   a. Client codes: Window size larger than 16 does not help; eight is good enough.
   b. Therapist codes: Window size 16 helps for difficult labels like Complex Reflections, but in general eight is good enough here too.

2. The impact of attention is mixed
   a. Word and sentence attention are not needed for categorizing client codes
   b. Both help for therapist codes

3. Paper also shows much more qualitative and quantitative error analysis
   a. Perhaps helpful for other dialogue modeling tasks too!

Take-away

Two new real-time dialogue observer tasks in therapy

Improvements from modeling innovations

Possible to predict, and give feedback on psychotherapy in real time (Tanana,

Thanks! Q & A?

Code : https://github.com/utahnlp/therapist-observer

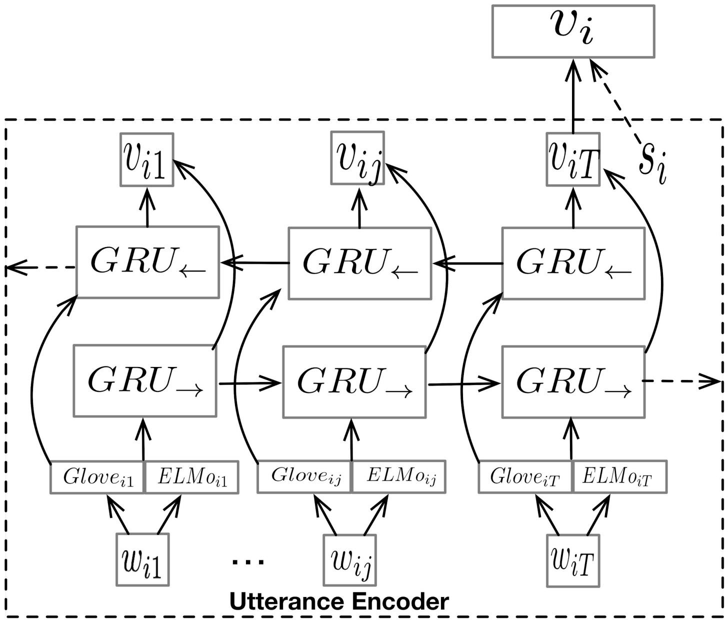
Extra slides

Here be dragons

Details of Hierarchical GRUs

Hierarchical GRU(HGRU)

Utterance Encoder (Bidirectional GRU)
Encoding **a sequence of words** in a sentence
Input: A sequence of word encoding vector
Output:
1. Task-specific contextualized word encoding
2. Utterance encoding vector



Hierarchical GRU(HGRU)

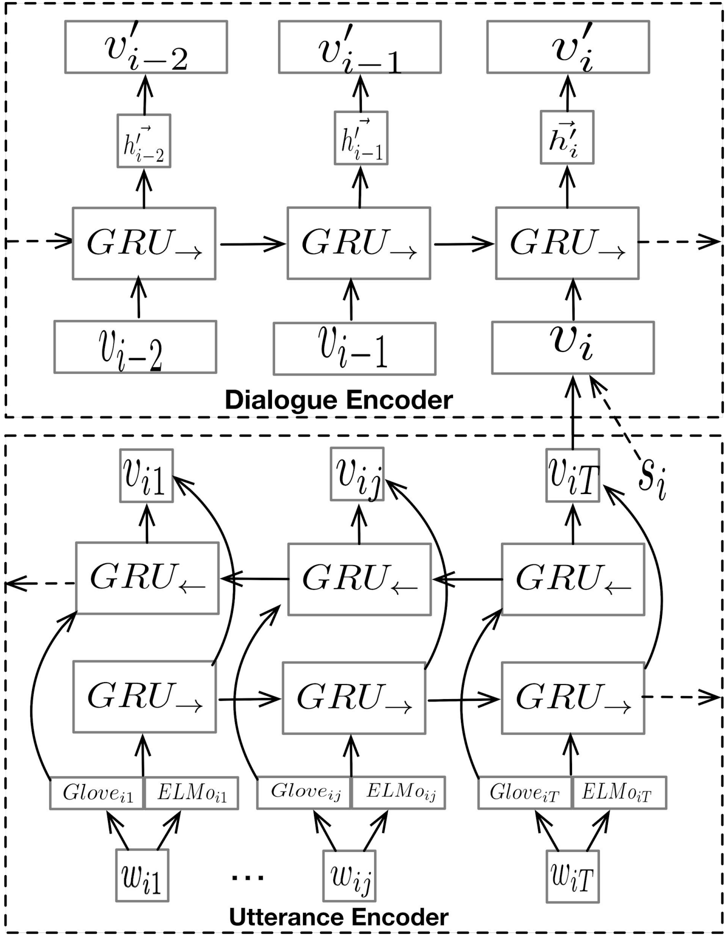
Dialogue Encoder (Uni-directional GRU)

Input: A sequence of utterance encoding vector
Output:
1. Task-specific contextualized utterance encoding
2. Dialogue encoding vector

Utterance Encoder (Bidirectional GRU)

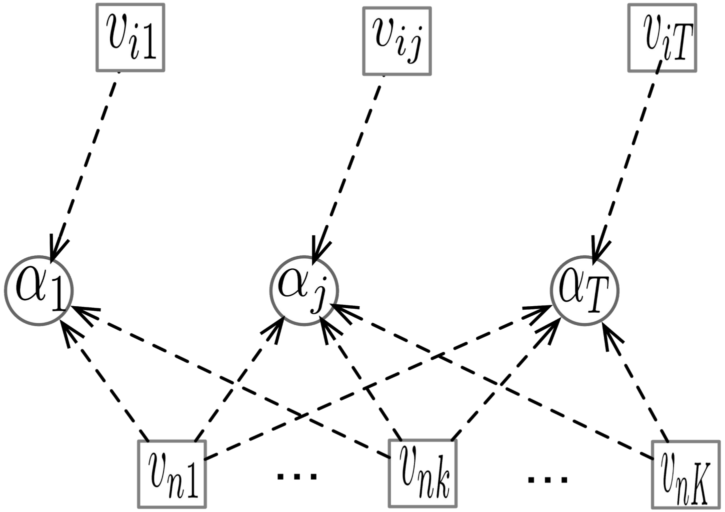
Input: A sequence of word encoding vector
Output:
1. Task-specific contextualized word encoding
2. Utterance encoding vector

HGRU, CONCAT



Word level attention: Details

Word-level Attention (Gated match-LSTM, BiDAF)



1. Match to get attention weight

$$\alpha_j^k = \frac{\exp(f_m(v_{nk}, v_{ij}))}{\sum_{j'} \exp(f_m(v_{nk}, v_{ij'}))}$$

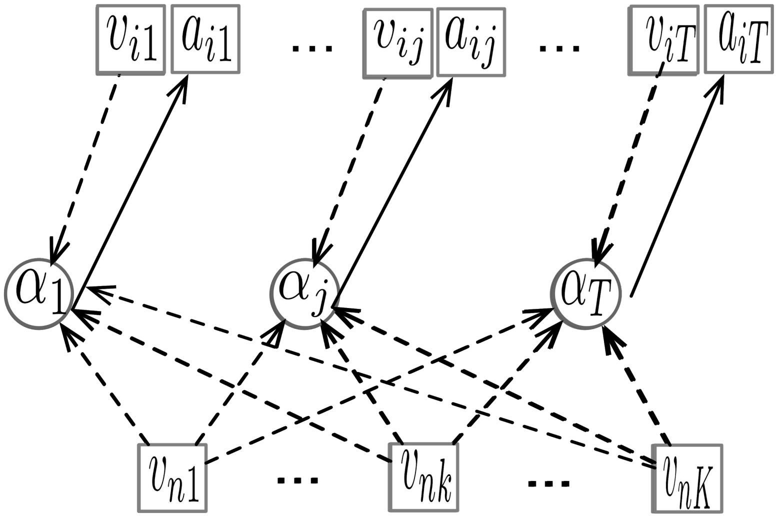
Word-level Attention (Gated match-LSTM, BiDAF)

2. Sum up useful info with attention weight

$$a_{ij} = \sum_k \alpha_j^k v_{nk}$$

1. Match to get attention weight

$$\alpha_j^k = \frac{\exp(f_m(v_{nk}, v_{ij}))}{\sum_{j'} \exp(f_m(v_{nk}, v_{ij'}))}$$



Word-level Attention (Gated match-LSTM, BiDAF)

3. Combine attended content with original content

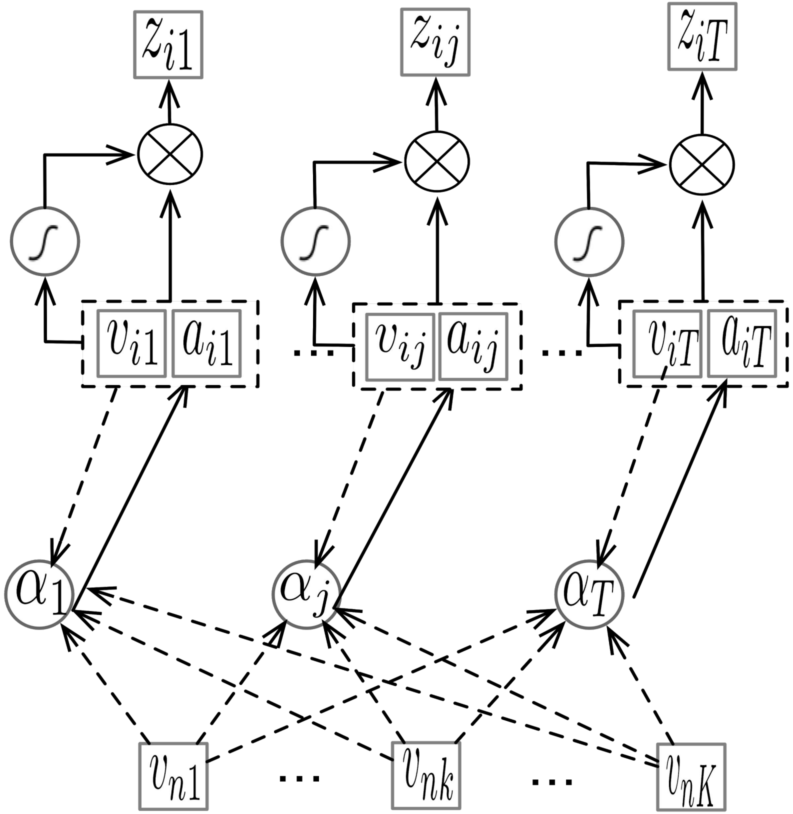
$$z_{ij} = f_c(v_{ij}, a_{ij})$$

2. Sum up useful info with attention weight

$$a_{ij} = \sum_k \alpha_j^k v_{nk}$$

1. Match to get attention weight

$$\alpha_j^k = \frac{\exp(f_m(v_{nk}, v_{ij}))}{\sum_{j'} \exp(f_m(v_{nk}, v_{ij'}))}$$

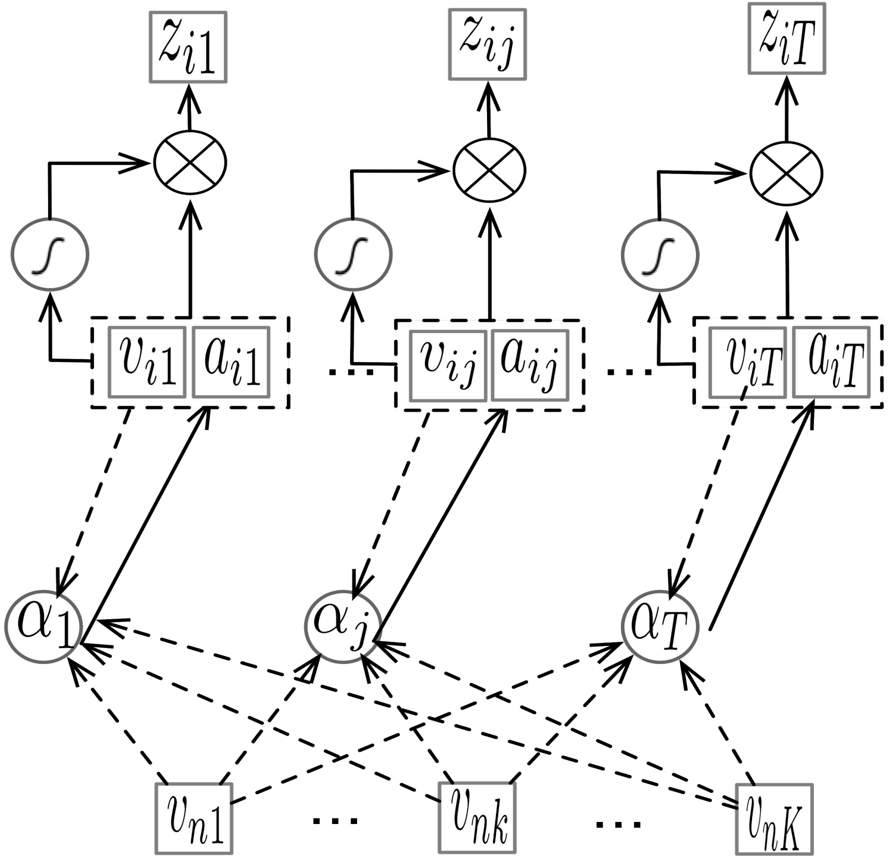


Word-level Attention (Gated match-LSTM, BiDAF)

By only adding two popular word-level attention mechanism **GMGRU** and **BiDAF** upon **HGRU**, we denote two models:

$$BiDAF^{H}$$

$$GMGRU^{H}$$



*In our experiments, we also tried word attention with **CONCAT, denoted as** $BiDAF^{C}$ $GMGRU^{C}$ **but not as good as hierarchical one in our tasks.**

Word-level Attention (Gated match-LSTM, BiDAF)

Method	f_m	f_c
BiDAF	$\boldsymbol{v}_{nk}\boldsymbol{v}_{ij}^T$	$[\boldsymbol{v}_{ij};\ \boldsymbol{a}_{ij};\ \boldsymbol{v}_{ij} \odot \boldsymbol{a}_{ij};\ \boldsymbol{v}_{ij} \odot \boldsymbol{a}']$
GMGRU	$\boldsymbol{w}^e \tanh(\boldsymbol{W}^k \boldsymbol{v}_{nk} + \boldsymbol{W}^q[\boldsymbol{v}_{ij}; \boldsymbol{h}_{j-1}])$	$[\boldsymbol{v}_{ij}; \boldsymbol{a}_{ij}]$

Two main subcomponent in attention:

1. Match function f_m
2. Combination function f_c

When only use word-level attention, we denote two models

$$BiDAF^H \qquad GMGRU^H$$

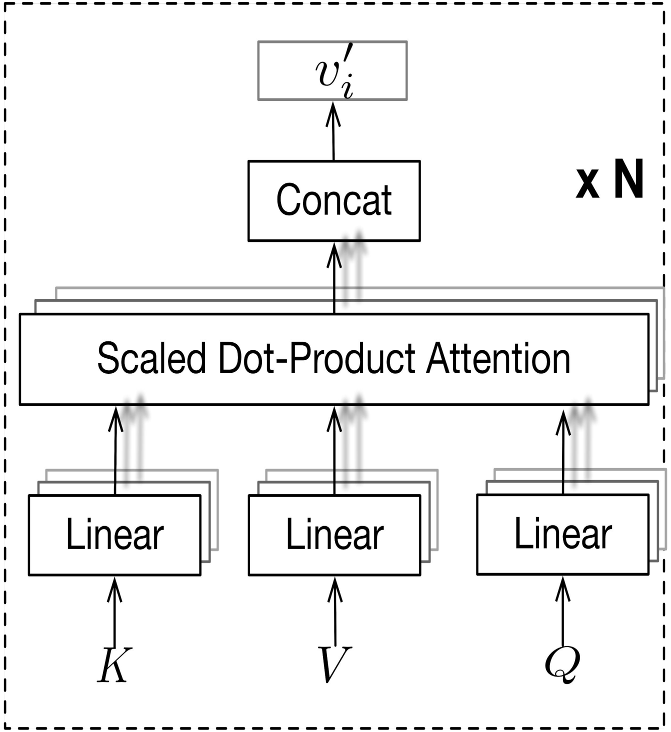
*In our experiments, we also tried word attention with **CONCAT, denoted as** $BiDAF^C$ $GMGRU^C$ **but not as good as hierarchical one in our tasks.**

Sentence-level Attention (Multi-head)

$$\text{Multihead}(Q, K, V) = [\text{head}_1; \cdots; \text{head}_h] W^O$$

$$\text{head}_i = \text{softmax}\left(\frac{QW_i^Q (KW_i^K)^T}{\sqrt{d_k}}\right) VW_i^V$$

Models	**Q**	**K = V**
$ANCHOR_{42}$	$[v_n]$	$[v_1 \cdots v_n]$
$SELF_{42}$	$[v_1 \cdots v_n]$	$[v_1 \cdots v_n]$



*We use **4 heads and N = 2 hops** for our transformer-based snt attention

References:

Gated match-LSTM:
Wenhui Wang, Nan Yang, Furu Wei, Baobao Chang, and Ming Zhou. 2017. **Gated self-matching networks for reading comprehension and question answering**. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), volume 1, pages 189–198

BiDAF:
Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. **Bidirectional attention flow for machine comprehension**. In ICLR.

Transformer Multihead attention:
Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. **Attention is all you need**. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, Advances in Neural Information Processing Systems 30, pages 5998–6008. Curran Associates, Inc.

Focal Loss:
Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. 2017. **Focal loss for dense object detection**. In Proceedings of the IEEE international conference on computer vision, pages 2980– 2988.

Results

Results Categorization

Best Categorization model for **client** is **HGRU**

any word or sentence attention we used didn't show extra improvements.

Method	macro	FN	CT	ST
Majority	30.6	**91.7**	0.0	0.0
Xiao et al. (2016)	50.0	87.9	32.8	29.3
$\text{BiGRU}_{\text{generic}}$	50.2	87.0	35.2	28.4
$\text{BiGRU}_{\text{ELMo}}$	52.9	87.6	**39.2**	32.0
Can et al. (2015)	44.0	91.0	20.0	21.0
Tanana et al. (2016)	48.3	89.0	29.0	27.0
CONCAT^{C}	51.8	86.5	38.8	30.2
GMGRU^{H}	52.6	89.5	37.1	31.1
BiDAF^{H}	50.4	87.6	36.5	27.1
$\mathcal{C}_C$	**53.9**	89.6	39.1	**33.1**
$\Delta = \mathcal{C}_C - \text{score}$	+3.5	-2.1	+3.9	+3.8

Best Categorization model for therapist:
use $GMGRU^{H}$ as word attention, $ANCHOR_{42}$ as sentence attention

Method	macro	FA	RES	REC	GI	QUC	QUO	MIA	MIN
Majority	5.87	47.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Xiao et al. (2016)	59.3	94.7	50.2	48.3	71.9	68.7	80.1	54.0	6.5
$\text{BiGRU}_{\text{generic}}$	60.2	94.5	50.5	49.3	72.0	70.7	80.1	54.0	10.8
$\text{BiGRU}_{\text{ELMo}}$	62.6	94.5	51.6	49.4	70.7	72.1	80.8	57.2	24.2
Can et al. (2015)	-	94.0	49.0	45.0	74.0	72.0	81.0	-	-
Tanana et al. (2016)	-	94.0	48.0	39.0	69.0	68.0	77.0	-	-
CONCAT^{C}	61.0	94.5	54.6	34.3	73.3	73.6	81.4	54.6	22.0
GMGRU^{H}	64.9	94.9	**56.0**	54.4	**75.5**	**75.7**	**83.0**	**58.2**	21.8
BiDAF^{H}	63.8	94.7	55.9	49.7	75.4	73.8	80.7	56.2	24.0
$\mathcal{C}_T$	**65.4**	**95.0**	55.7	**54.9**	74.2	74.8	82.6	56.6	**29.7**
$\Delta = \mathcal{C}_T - \text{score}$	+5.2	+0.3	+3.9	+3.8	+0.2	+2.8	+1.6	+2.6	+18.9

Results
Forecasting

Best forecasting model for client and therapist: $SELF_{42}$

Method	Recall	F1								
	R@3	macro	FA	RES	REC	GI	QUC	QUO	MIA	MIN
CONCAT^F	72.5	23.5	63.5	0.6	0.0	53.7	27.0	15.0	18.2	9.0
HGRU	76.0	28.6	71.4	12.7	**24.9**	58.3	28.8	5.9	**17.4**	9.7
GMGRU^H	76.6	26.6	**72.6**	10.2	20.6	58.8	27.4	6.0	8.9	7.9
$\mathcal{F}_T$	**77.0**	**31.1**	71.9	**19.5**	24.7	**59.2**	**29.1**	**16.4**	15.2	**12.8**

Ablation Study on Categorizing Client Codes

Our selected model are **HGRU**

Ablation	Options	macro	FN	CT	ST
history window size	0	51.6	87.6	39.2	32.0
	4	52.6	88.5	37.8	31.5
	8*	53.9	89.6	39.1	33.1
	16	52.0	89.6	39.1	33.1
word attention	+ GMGRU	52.6	89.5	37.1	31.1
	+ BiDAF	50.4	87.6	36.5	27.1
sentence attention	+ SELF_{42}	53.9	89.2	39.1	33.2
	+ ANCHOR_{42}	53.0	88.2	38.9	32.0

1. Context helps for categorizing client codes; Window size larger than 16 does not help for client code
2. Word Attention generally does not help for categorizing client codes
3. Sentence Attention generally does not help for categorizing client codes

Ablation Study on Categorizing Therapist Codes

Our selected model are $GMGRU^{H}$ + $ANCHOR_{42}$

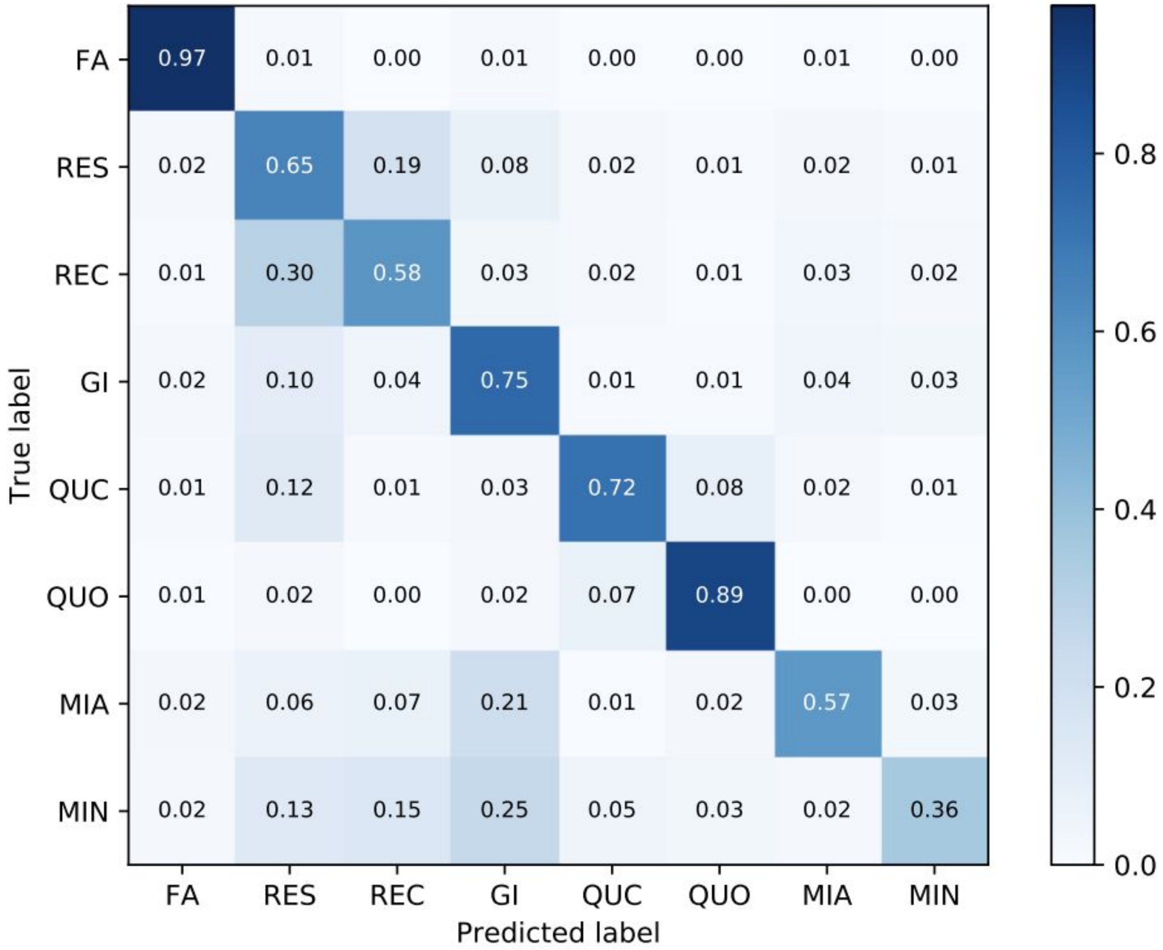
Ablation	Options	macro	RES	REC	MIN
history window size	0	62.6	51.6	49.4	24.2
	4	64.4	54.3	53.2	23.7
	8*	65.4	55.7	54.9	29.7
	16	**65.6**	55.4	**56.7**	26.7
word attention	- GMGRU	62.0	51.9	51.7	16.0
	\ BiDAF	63.5	54.2	51.3	22.6
sentence attention	- $ANCHOR_{42}$	64.9	56.0	54.4	21.8
	\ $SELF_{42}$	63.4	55.5	48.2	21.1

1. Larger context size can even help, especially for REC
2. Adding Word Attention generally helps for categorizing therapist code; GMGRU helps more than BiDAF
3. ANCHOR Based sentence attention performs better than Self-attention in our case.

Error breakdown for categorizing client codes

Category and Explaination	Client Examples (Gold MISC)
Reasoning is required to understand whether a client wants to change behavior, even with full context (50,42)	T: On a scale of zero to ten how confident are you that you can implement this change ? C: I don't know, seven maybe (CT); I have to wind down after work (ST)
Concise utterances which are easy for humans to understand, but missing information such as coreference, zero pronouns (22,31)	I mean I could try it (CT) Not a negative consequence for me (ST) I want to get every single second and minute out of it(CT)
Extremely short ($\leq$ 5) or long sentence ($\geq$ 40), caused by incorrect turn segementation. (21,23)	It is a good thing (ST) Painful (CT)
Ambivalent speech, very hard to understand even for human. (7,4)	What if it does n't work I mean what if I can't do it (ST) But I can stop whenever I want(ST)

Confusion Matrix for categorizing therapist codes



Impact of Focal Loss

Loss	Client			Therapist				
	F_1	CT	ST	F_1	RES	REC	MIA	MIN
$\mathcal{C}^{\text{ce}}$	47.0	28.4	22.0	60.9	54.3	53.8	53.7	4.8
$\mathcal{C}^{\text{wce}}$	53.5	39.2	32.0	65.4	55.7	54.9	56.6	29.7
$\mathcal{C}^{\text{fl}}$	53.9	39.1	33.1	65.4	55.7	54.9	56.6	29.7
$\mathcal{F}^{\text{ce}}$	42.1	17.7	18.5	26.8	3.3	20.8	16.3	8.3
$\mathcal{F}^{\text{wce}}$	43.1	20.6	23.3	30.7	17.9	25.0	17.7	10.9
$\mathcal{F}^{\text{fl}}$	44.2	24.7	22.7	31.1	19.5	24.7	15.2	12.8

$\gamma = 1$

$\gamma = 0$

$\gamma = 1$

$\gamma = 3$

We choose to balance weights as {1.0,1.0,0.25} for CT,ST and FN respectively
and {0.5, 1.0, 1.0, 1.0, 0.75, 0.75,1.0,1.0} for FA, RES, REC, GI, QUC, QUO, MIA, MIN

- Focal loss helps most for categorizing client codes.
- It also slightly helps when comparing to weighted cross entropy for other models.