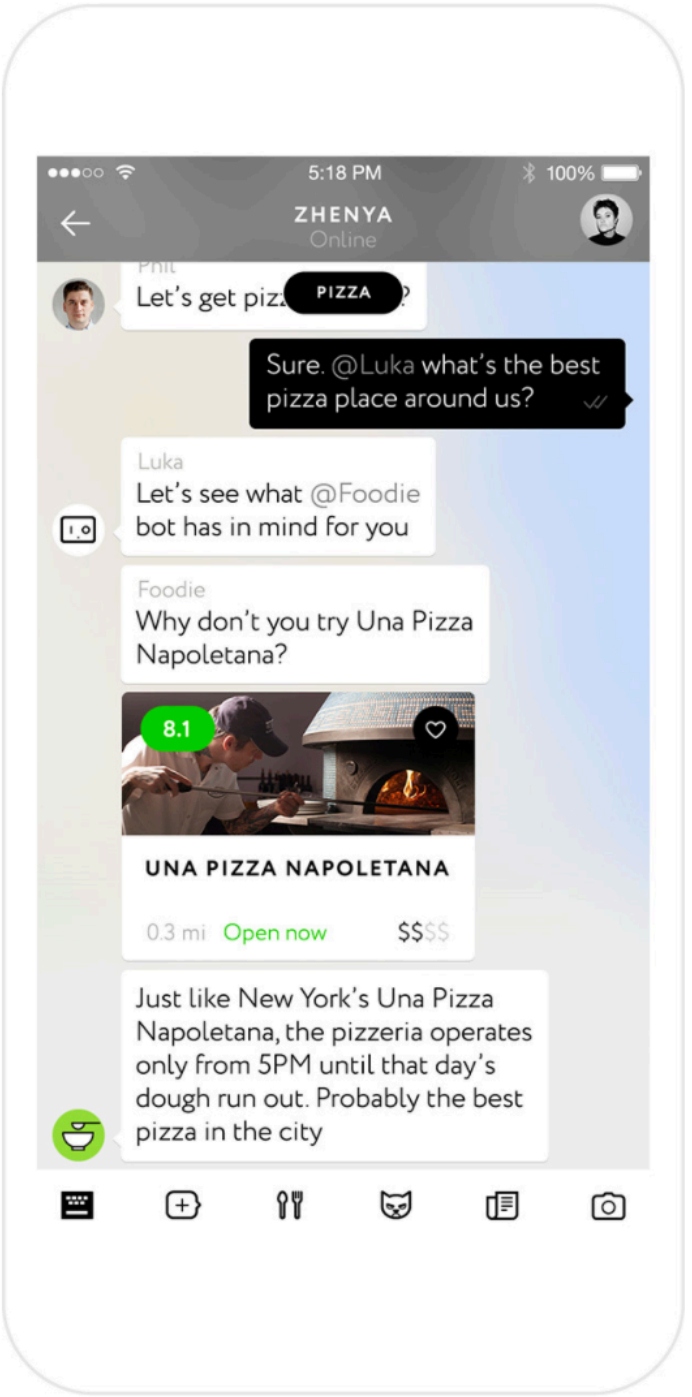
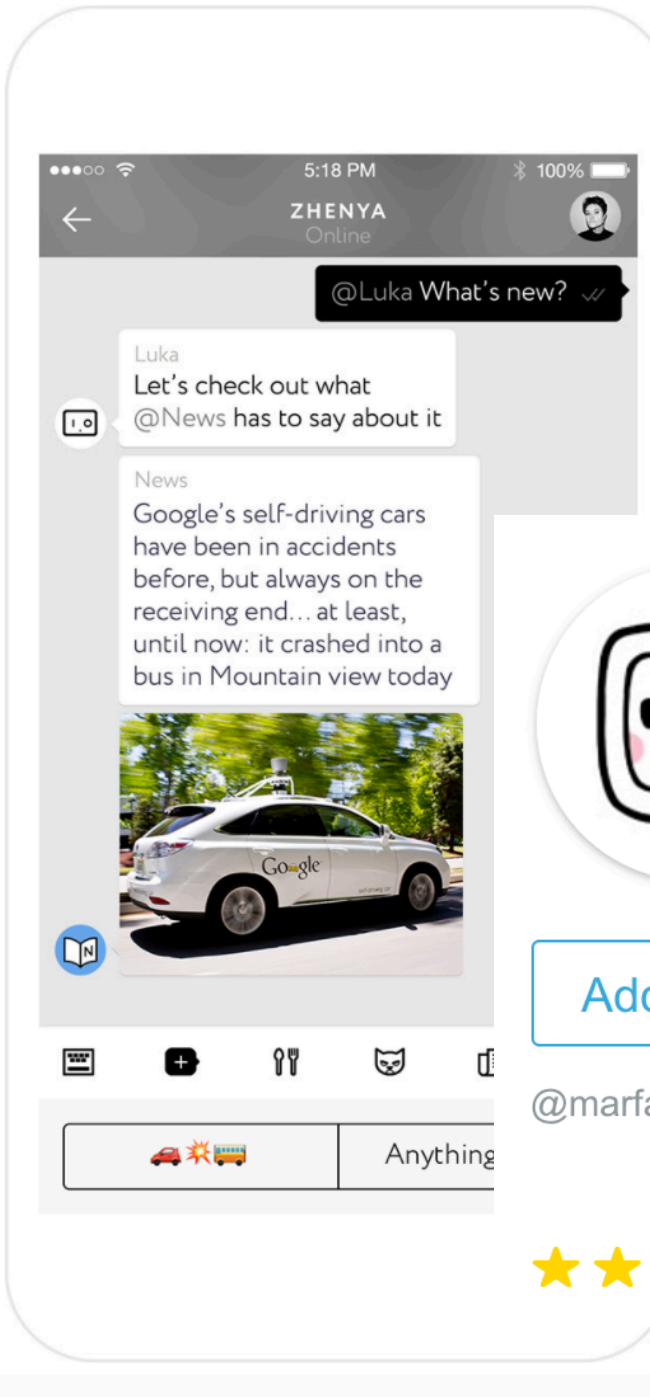
Neural Conversational Models

Michael Khalman, Luka Labs

Goal of this talk

- How to build a simple neural chatbot
- How small-talk chatbots can be applied in a real product
- What are the main research directions in this area





24 мая 2016, 19:00 Вадим Елистратов 13 744 27

Стартап Luka создал чат-бота на основе личных сообщений и записей в соцсетях погибшего

Стартап Luka выпустил для своего мессенджера чат-бота, воспроизводящего манеру общения погибшего в ДТП Романа Мазуренко — бывшего арт-директора «Стрелки» и основателя стартапа Stampsy. Об этом на Фейсбуке сообщила генеральный директор Luka Евгения Куйда.

Add To

@marfa_bot

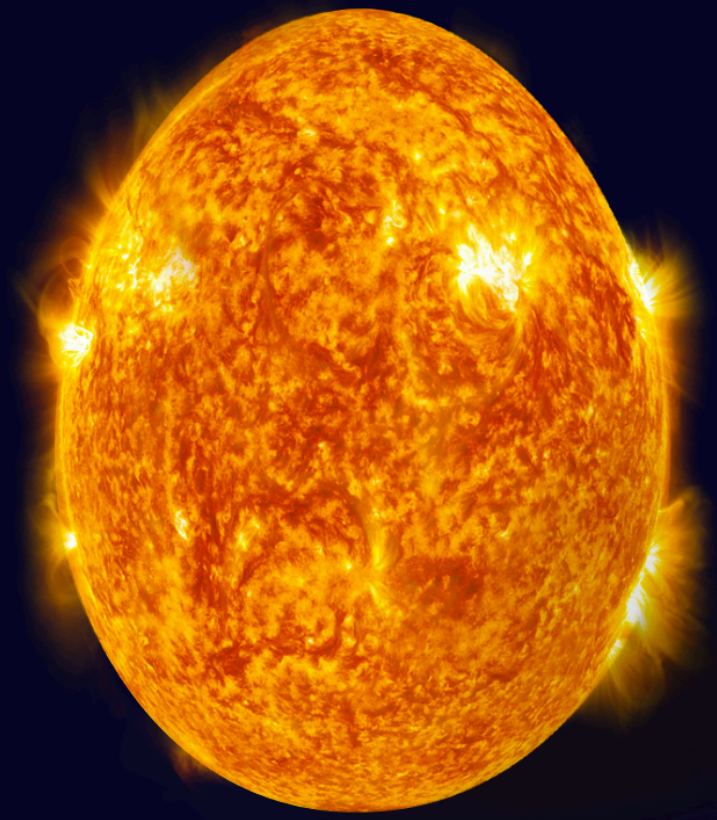
★★★★☆ (639)

Meet Luka, a messenger for bots and humans

Some facts about people/teenagers

- People are lonely
- People have **a lot** of free time
- People love games

Replika



Hi, I am your radiant Replika. What is my name?

Enter your Replika's name

Reserve Now

Replika is your AI friend that you teach and grow through conversation. Reserve your name now and be the first to start raising your Replika when the app is out!

Coming soon on the App Store

Our dialogue architecture

- Hand-crafted scenarios
- Selective model
- **Generative model**

Small-talk chatbots	Goal-based systems
Applications	
Entertainment	Personal Assistant
Responses	
• Somewhat relevant • Context-aware • **Interesting/Diverse** • **Stimulate conversation**	• **Relevant** • Context-aware
Metrics	
• A/B tests • Assessors	• Precision/Recall • Accuracy

Outline

1. Language Modelling
2. Basic neural conversational models
3. How to generate samples?
4. Advanced neural conversational models

Outline

1. **Language Modelling**
2. Basic neural conversational models
3. How to generate samples?
4. Advanced neural conversational models

Language Modelling

Let $w_1, w_2, \dots, w_n$ – sequence of words

$p(w_1, w_2, \dots, w_n)$ – ?

Language Modelling

Let $w_1, w_2, \dots, w_n$ – sequence of words

$$p(w_1, w_2, \dots, w_n) = \prod_{i=1}^{n} p(w_i | w_{i-1}, \dots, w_1)$$

Language Modelling: n-gram models

Let $w_1, w_2, \dots, w_n$ – sequence of words

$$p(w_1, w_2, \dots, w_n) \approx \prod_{i=1}^{n} p(w_i | w_{i-k}, \dots, w_{i-1})$$

$p(w_i | w_{i-k}, \dots, w_{i-1}) =$

"what is the probability of seeing w_i after $w_{i-k}, \dots, w_{i-1}$"

Language Modelling: n-gram models

$$p(w_1, w_2, \dots, w_n) \approx \prod_{i=1}^{n} p(w_i | w_{i-k}, \dots, w_{i-1})$$

How to train?

Just compute and renormalize counts!

$\text{count}(w_1, \dots, w_k) = \text{“how many times}$
$\text{n}-\text{gram } (w_1, \dots, w_k) \text{ occur in text”}$

Language Modelling: n-gram models

$$p(w_1, w_2, \ldots, w_n) \approx \prod_{i=1}^{n} p(w_i | w_{i-k}, \ldots, w_{i-1})$$

Bigger $k =>$ more accurate/overfitted model,
but memory-requirements scale **exponentially** with k
Many ways to get stable and smoothed estimated:

- Dirichilet prior
- Back-offs
- Kneser-Ney smoothing

Language Modelling: n-gram models

\+

- Really simple model
- Works surprisingly well

–

- Needs a lot of memory
- Limited context size

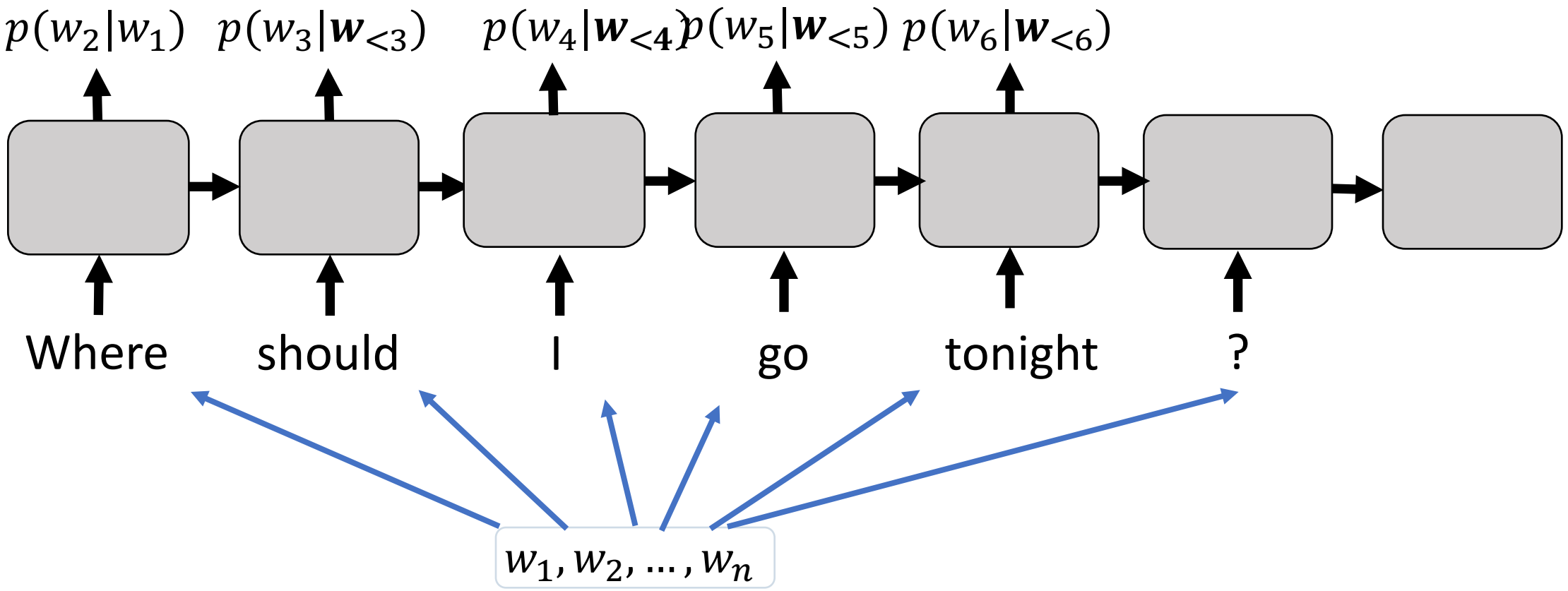
Language Modelling: continuous space language models

$$p(w_1, w_2, \dots, w_n) = \prod_{i=1}^{n} p(w_i | w_1, \dots, w_{i-1})$$

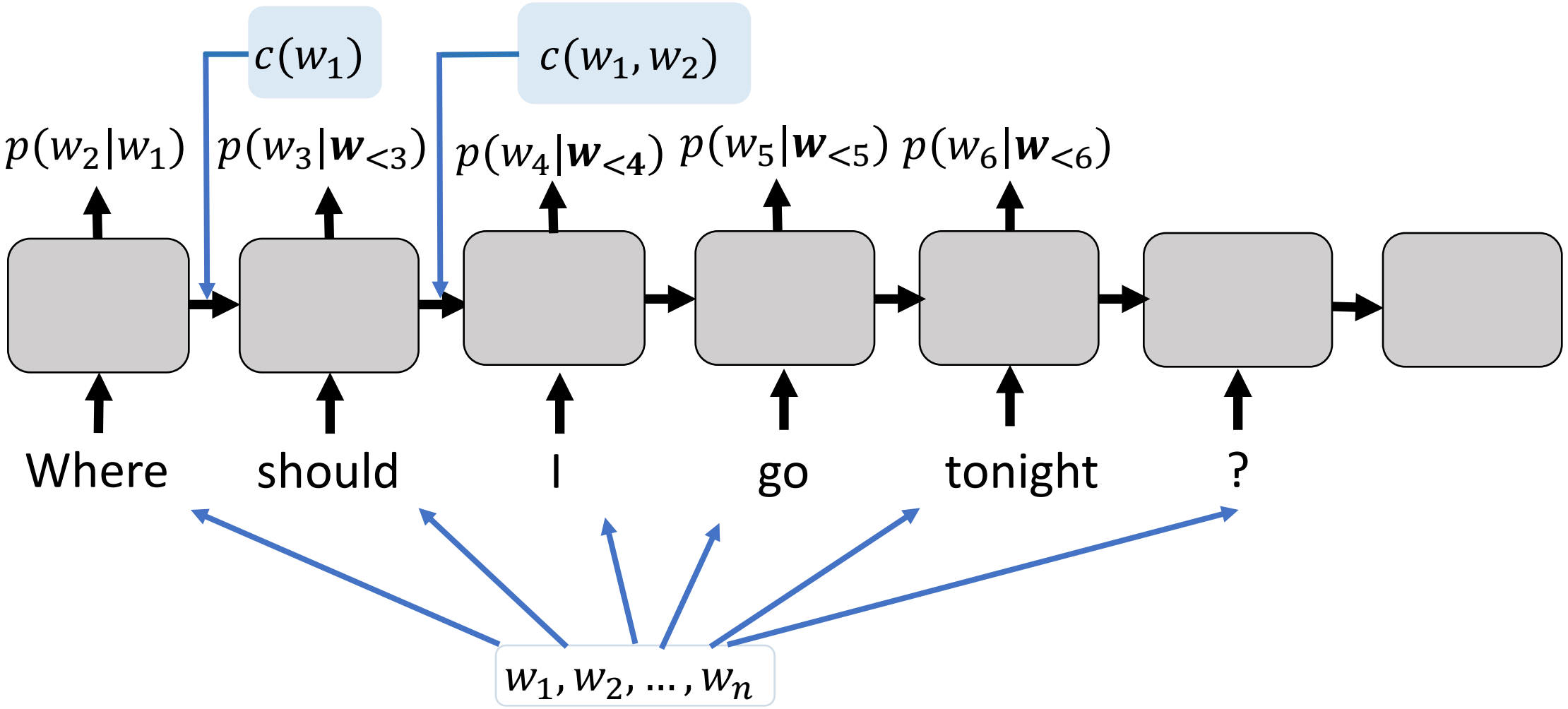
$$p(w_i | w_1, \dots, w_{i-1}) \approx f(w_i | c(w_1, \dots, w_{i-1}))$$

$f(\dots)$ and $c(\dots)$ are neural networks, for example an RNN/LSTM

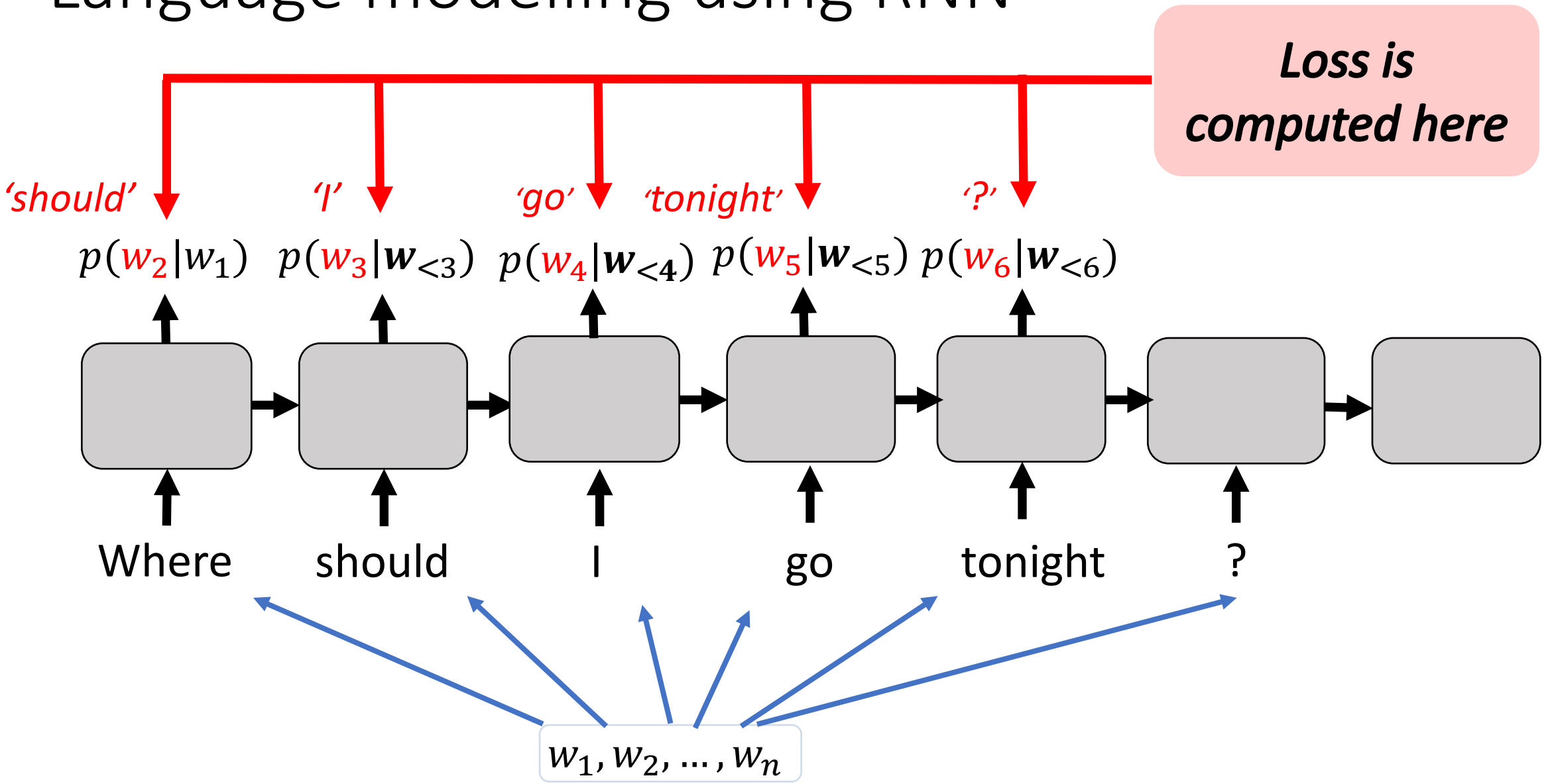
Language modelling using RNN



Language modelling using RNN



Language modelling using RNN



Language modelling using RNN

+

- Captures complicated dependencies for (in theory) arbitrary large lag
- Scales well on big corpora
- Can be extended to more complicated use-cases (Im2Caption, MT, speech recongnition, **chatbots** etc)

–

- Training of RNN needs a lot of hacks/GPU/time

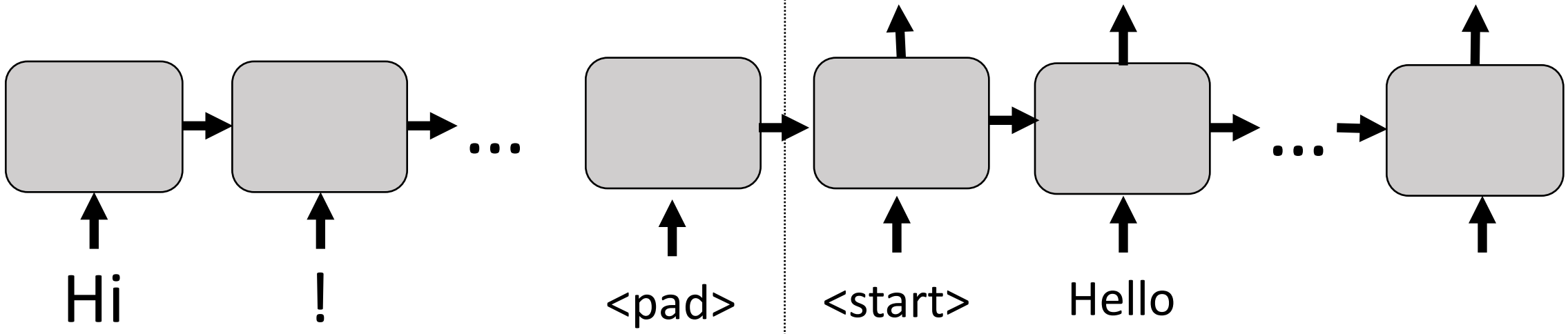
Outline

1. Language Modelling
2. **Basic neural conversational models**
3. How to generate samples?
4. Advanced neural conversational models

Basic conversational model

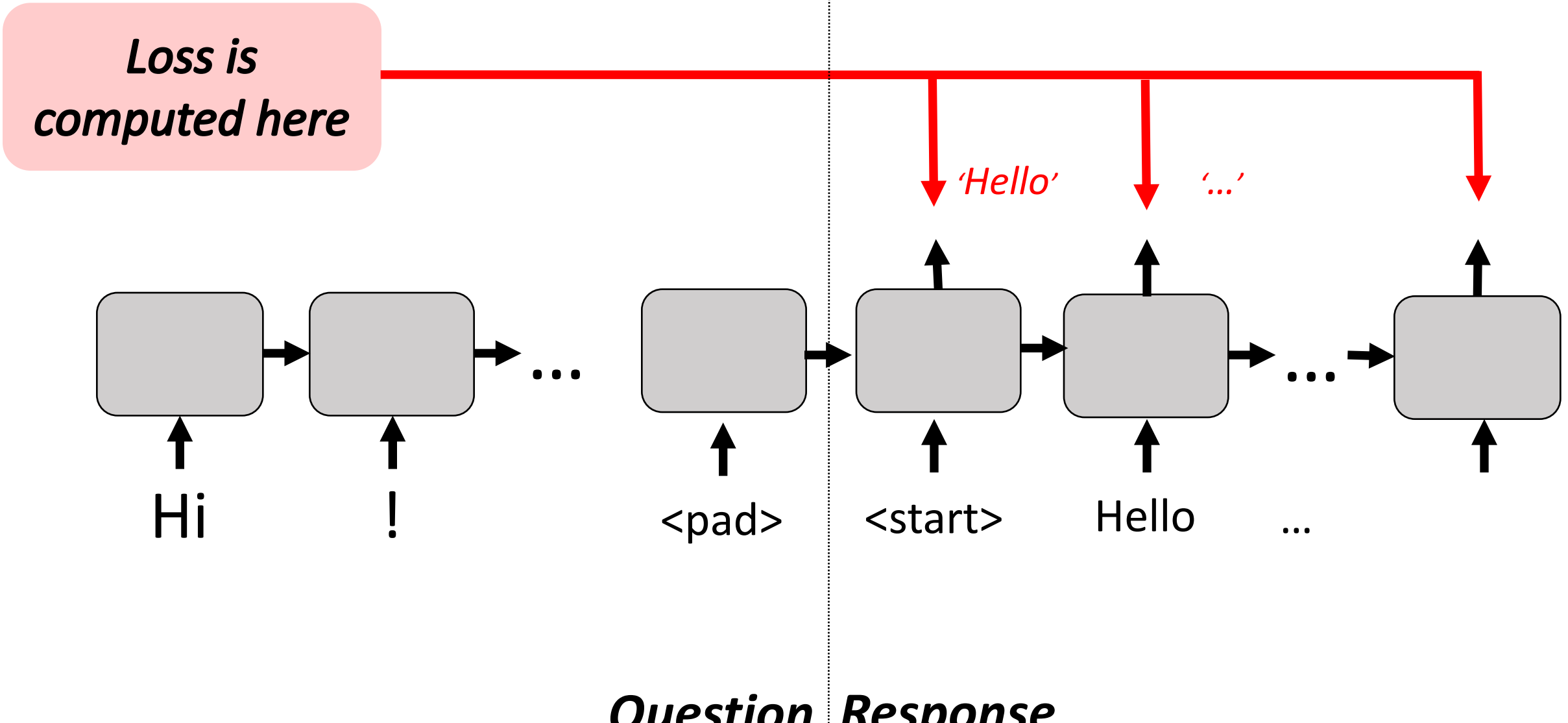
How to turn your language model into a conversational model?

Basic conversational model

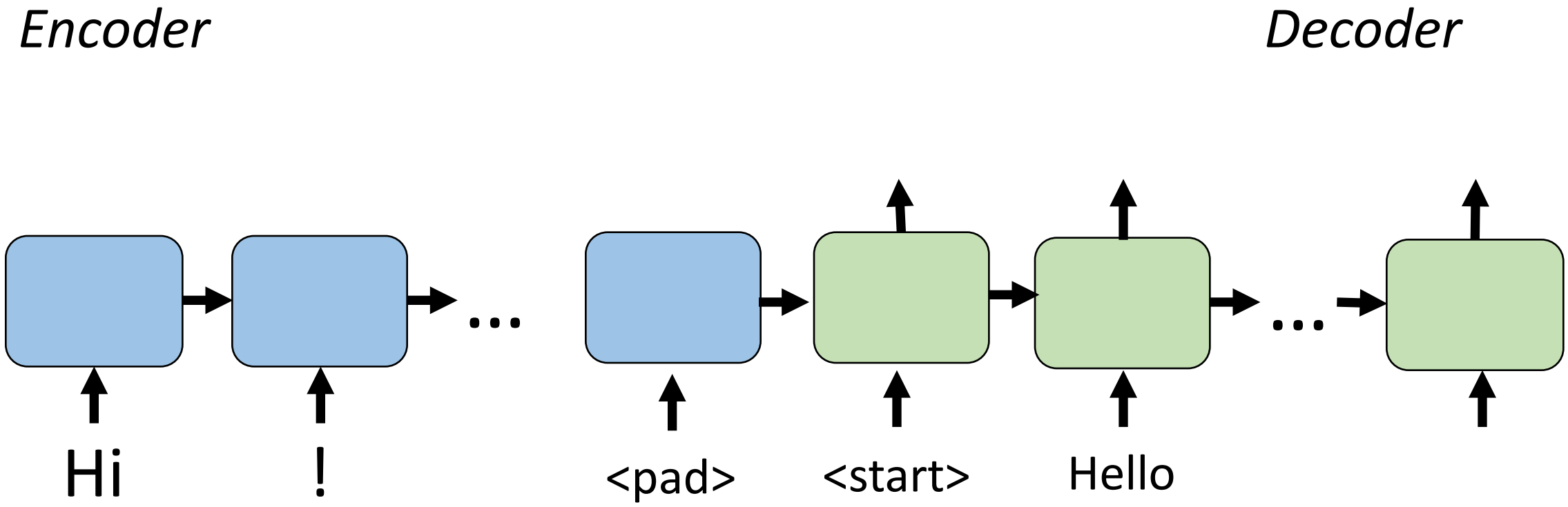


Question ***Response***

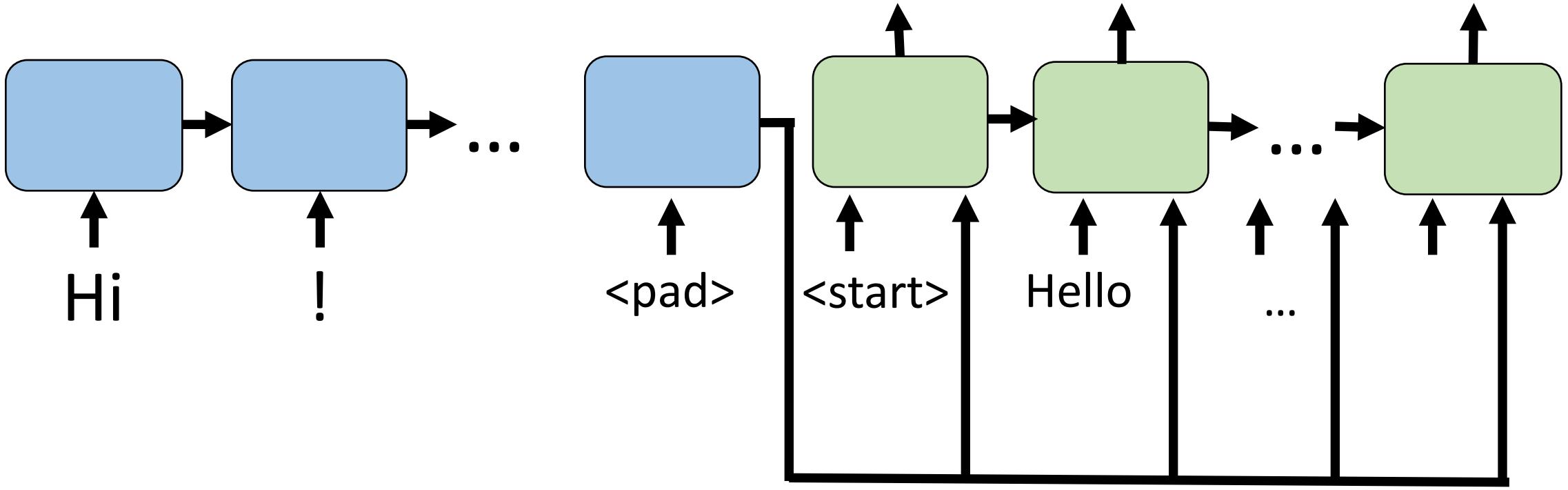
Basic conversational model



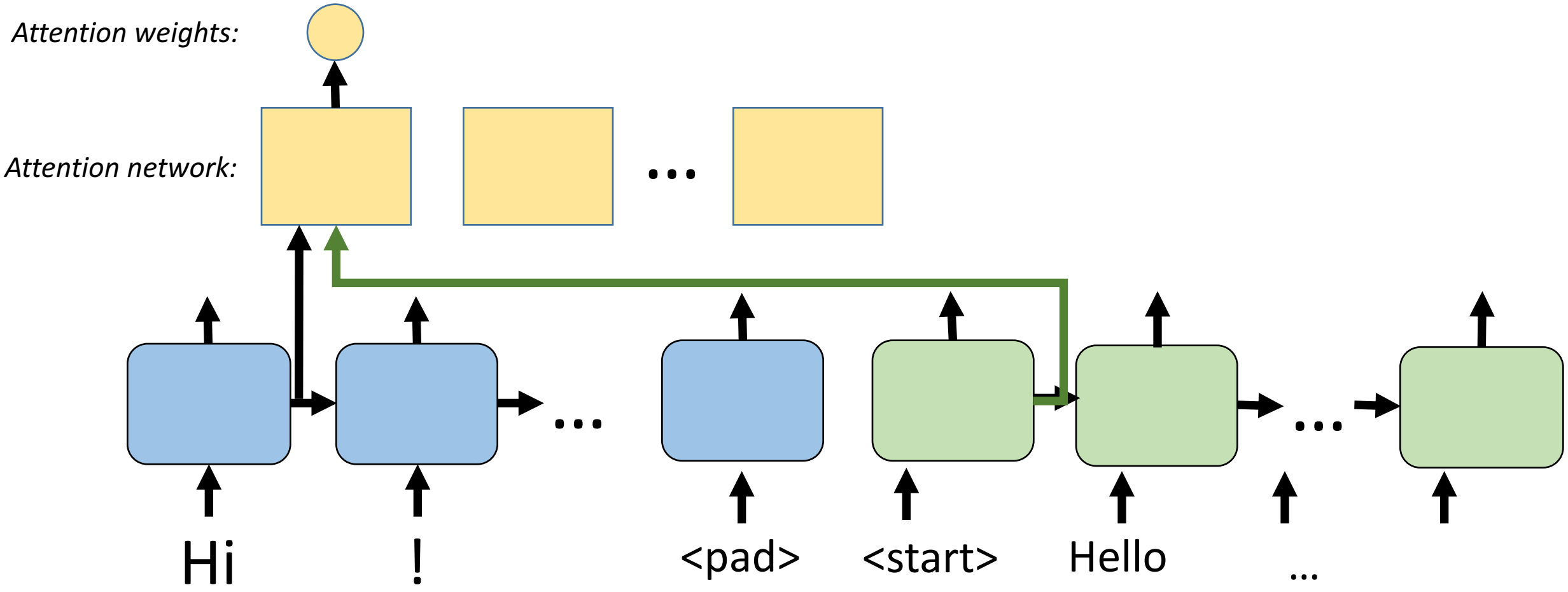
Sequence to sequence: seq2seq



Sequence to sequence: seq2seq

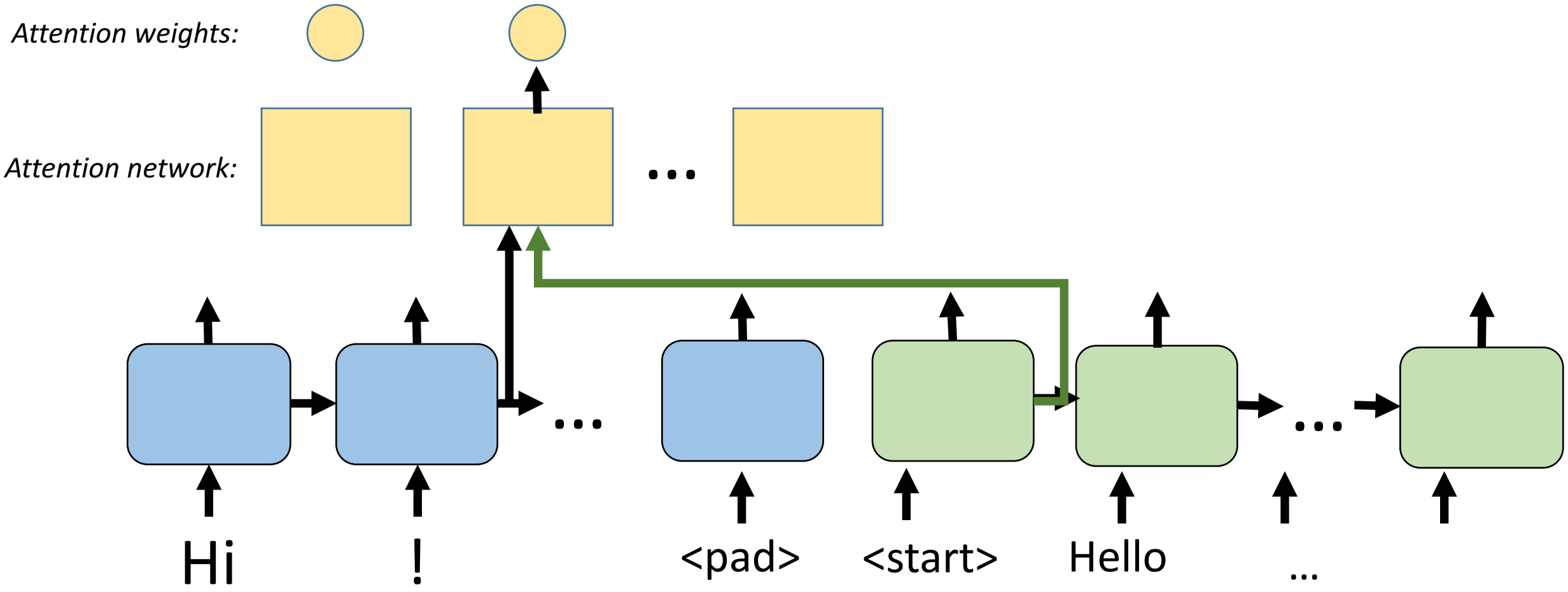


Sequence to sequence: seq2seq + attention



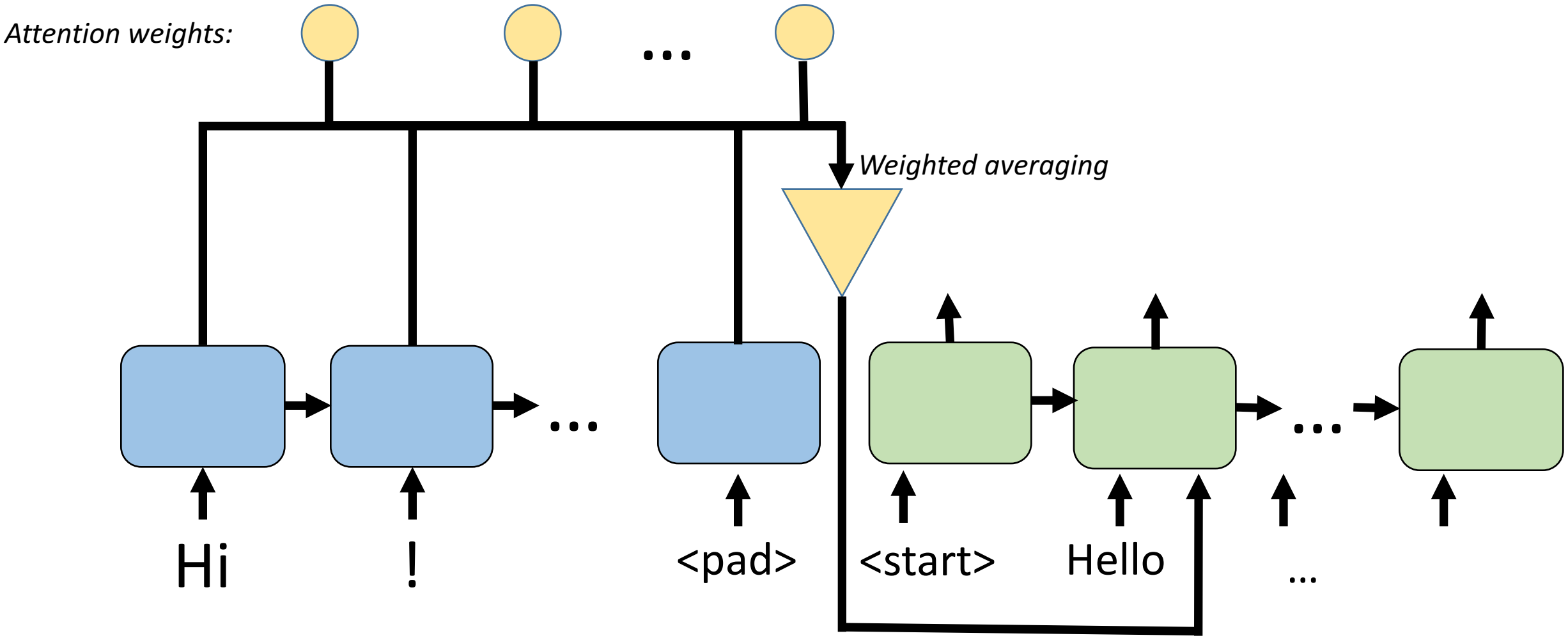
Bahdanau D., Cho K., Bengio Y. Neural machine translation by jointly learning to align and translate

Sequence to sequence: seq2seq + attention



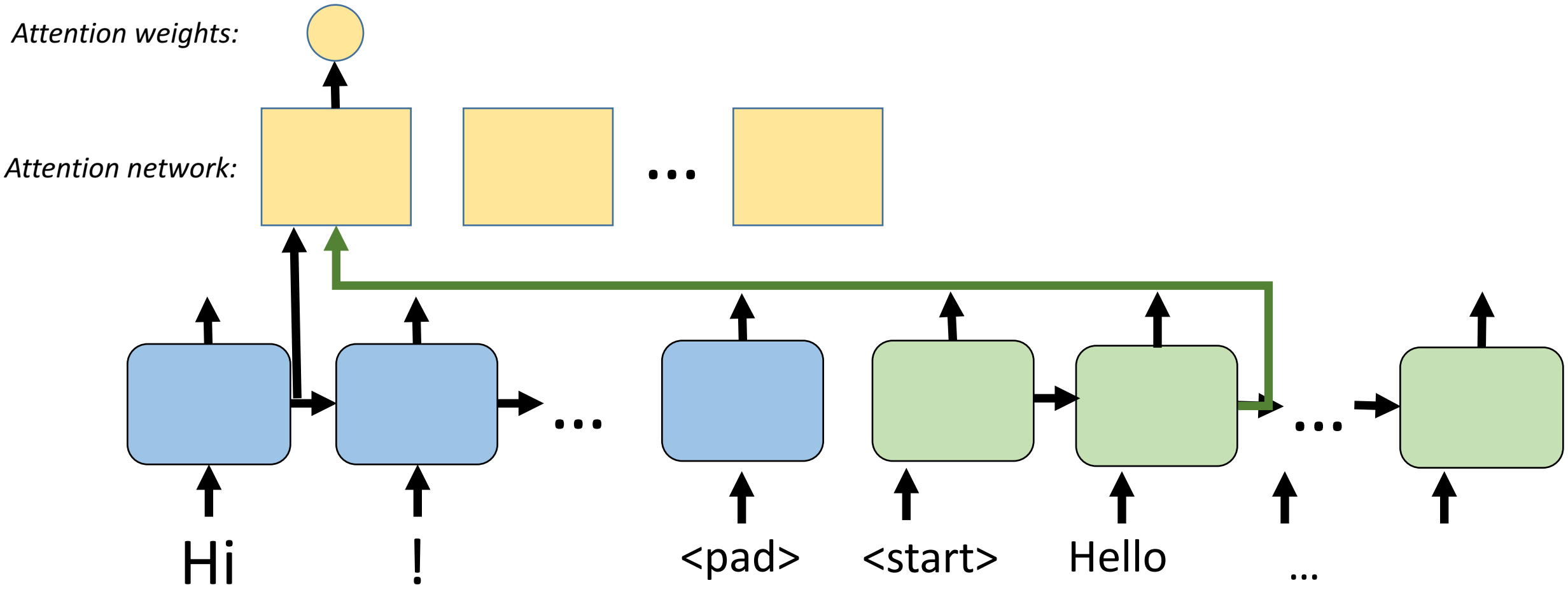
Bahdanau D., Cho K., Bengio Y. Neural machine translation by jointly learning to align and translate

Sequence to sequence: seq2seq + attention



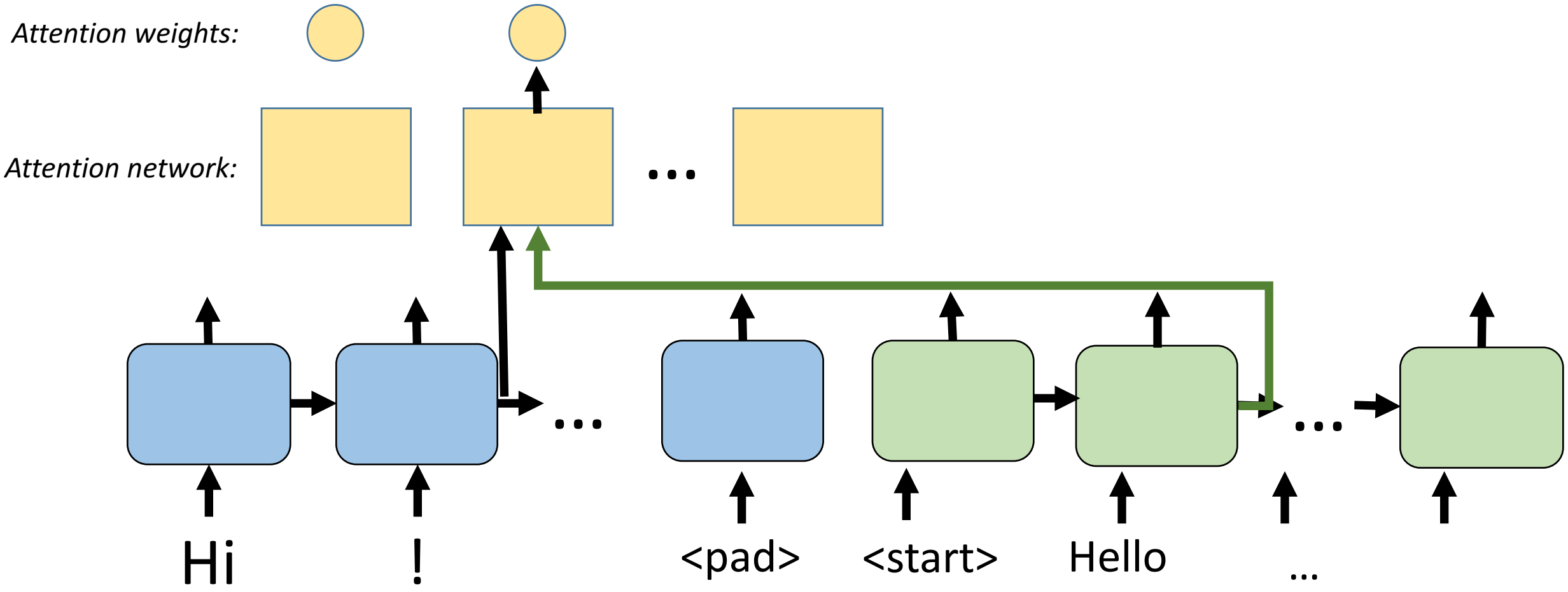
Bahdanau D., Cho K., Bengio Y. Neural machine translation by jointly learning to align and translate

Sequence to sequence: seq2seq + attention



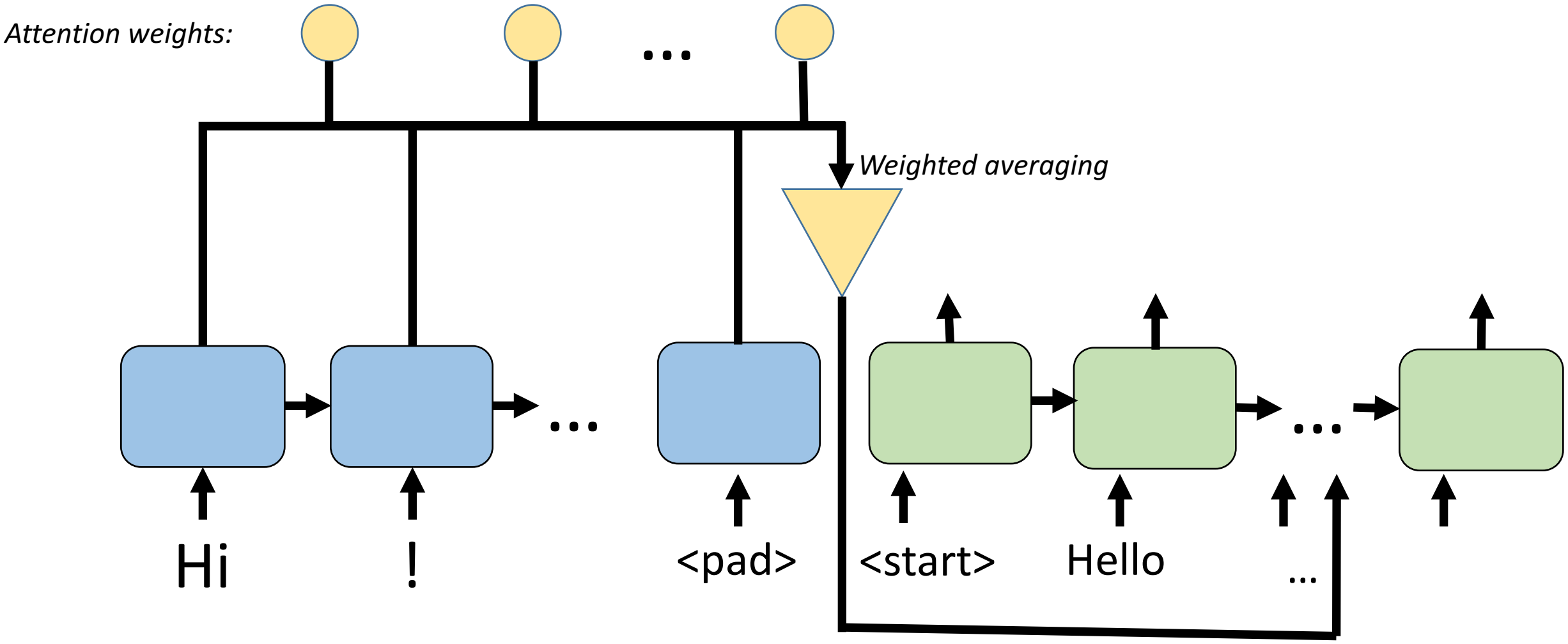
Bahdanau D., Cho K., Bengio Y. Neural machine translation by jointly learning to align and translate

Sequence to sequence: seq2seq + attention



Bahdanau D., Cho K., Bengio Y. Neural machine translation by jointly learning to align and translate

Sequence to sequence: seq2seq + attention



Bahdanau D., Cho K., Bengio Y. Neural machine translation by jointly learning to align and translate

Datasets

- Twitter
- Open subtitles
- Reddit comments
- …

Outline

1. Language Modelling
2. Basic neural conversational models
3. **How to generate samples?**
4. Advanced neural conversational models

Sequence to sequence: how to generate samples?

So we have our distribution $p(y \mid x)$

How to generate the response y to a question x?

Sequence to sequence: how to generate samples?

So we have our distribution $p(y \mid x)$

How to generate the response y to a question x?

1. Sample from $p(y \mid x)$?

Sequence to sequence: how to generate samples?

I trained a seq2seq on 37m twitter dialogues (~9GB of raw text) for 4 days and tried to have a conversation with it.

Sequence to sequence: example (sampling)

- **Me: Hello**
- **Bot: Hello, cool. ^. ^**
- **Me: What should I do tonight?**
- **Bot: By my email... Ill ask your brother!**
- **Me: But I don't have a brother**
- **Bot: The way you do, and pamela * you don't leave her in a corner]**
- **Me: What are you talking about?**
- **Bot: White developers**
- **Me: Why white?**
- **Bot: Geez. U just need the full oil. Lmao.**
- **Me: You are so useless…**
- **Bot: I will have to kick in some heavy pieces. Silly bruhh!!**

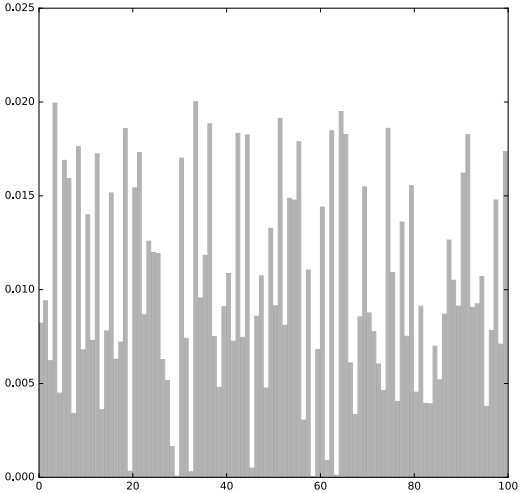
Sampling with temperature

Let $p(y \mid x) = e^{-E(y)}$, then sample from $\tilde{p}(y|x) = e^{-\frac{E(y)}{T}}$,

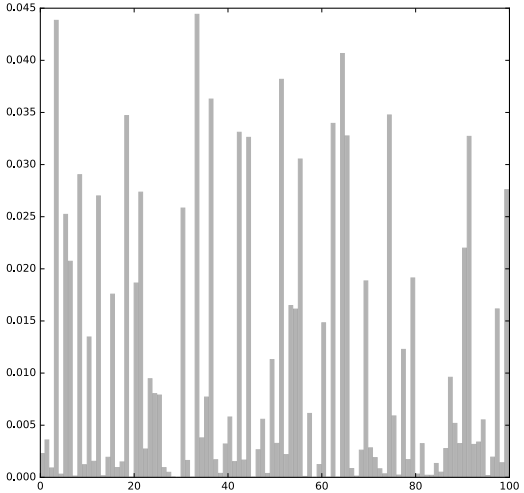
where $T > 0$ – temperature.

When $T < 1$ the distribution gets **sharper**.

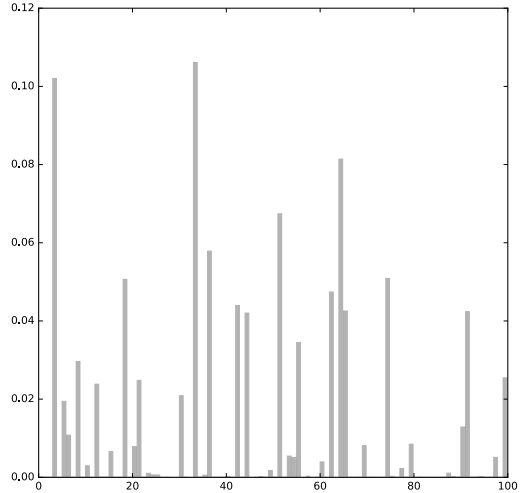
T = 1



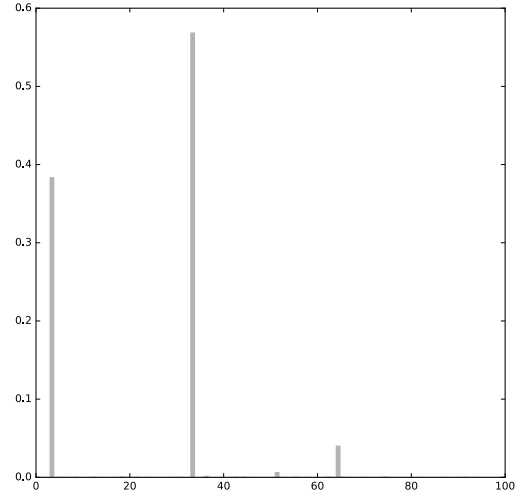
T = 0.3



T = 0.1



T = 0.01



Sequence to sequence: example (sampling, $T = 0.5$)

- **Me: Hello**
- **Bot: Hello, Ivory. Make it quick!**
- **Me: What should I do tonight?**
- **Bot: You should come to my house**
- **Me: Do you really think so? What should I wear?**
- **Bot: I'm not sure but I'm going to try to find a way to get a job**
- **Me: What do you like more, black jacket or white t-shirt?**
- **Bot: There's a bunch of people who don't want to talk about it.**
- **Me: About what?**
- **Bot: Like a whole bunch of people.**
- **Me: Sorry, I don't get it :(**
- **Bot: It's ok. I'm sure you'll have a great time**

Sequence to sequence: example (sampling, $T = 0.2$)

- **Me: Hello**
- **Bot: Hello…**
- **Me: What should I do tonight?**
- **Bot: I'm not sure**
- **Me: Okay, I'll go to the party. What should I wear?**
- **Bot: I'm wearing a dress.**
- **Me: What do you like more, black jacket or white t-shirt?**
- **Bot: I like the red shirt.**
- **Me: But I don't have a red shirt!**
- **Bot: I'm sorry…**
- **Me: Can I borrow your shirt?**
- **Bot: I don't have a shirt.**

Sampling with temperature

Sampling with temperature $T \to 0$ leads to greedy selecting the next word.

That’s not a good idea usually.

Sequence to sequence: how to generate samples?

So we have our distribution $p(y \mid x)$

How to generate the response y to a question x?

1. Sample from $p(y \mid x)$ with temperature.

Sequence to sequence: how to generate samples?

So we have our distribution $p(y \mid x)$

How to generate the response y to a question x?

1. Sample from $p(y \mid x)$ with temperature.
2. Find approximate mode?

Sequence to sequence: example (mode)

- **Me: Hello**
- **Bot: Hello**
- **Me: What should I do tonight?**
- **Bot: Text me**
- **Me: Do you really think so? What should I text?**
- **Bot: I don't know.**
- **Me: What do you like more, black jacket or white t-shirt?**
- **Bot: Black.**
- **Me: Do you mean black jacket?**
- **Bot: Yeah**
- **Me: Okay then. I'll text you in a black jacket**
- **Bot: Okay.**

MMI: Maximum Mutual Information criterion

Instead of $\hat{y} = \underset{y}{\operatorname{argmax}}\, p(y \mid x)$, find:

$$\hat{y} = \underset{y}{\operatorname{argmax}} \frac{p(y,x)}{p(y)p(x)} = \underset{y}{\operatorname{argmax}} \left[\log p(y|x) - \log p(y)\right]$$

Li J. et al. A diversity-promoting objective function for neural conversation models

MMI: Maximum Mutual Information criterion

$$\hat{y} = \underset{\mathrm{y}}{\operatorname{argmax}} \left[\log \mathrm{p}(\mathrm{y}|\mathrm{x}) - \log \mathrm{p}(\mathrm{y})\right]$$

$p(y)$ can be modeled as a RNN language model.

MMI: Maximum Mutual Information criterion

$$\hat{y} = \underset{\mathrm{y}}{\operatorname{argmax}} \left[\log \mathrm{p}(\mathrm{y}|\mathrm{x}) - \lambda \log \mathrm{p}(\mathrm{y}) \right]$$

$p(y)$ can be modeled as a RNN language model.

MMI: Maximum Mutual Information criterion

$$\hat{y} = \underset{y}{\operatorname{argmax}} \left[\log p(y|x) - \lambda \log p(y) \right]$$

$p(y)$ can be modeled as a RNN language model.

Apply Bayes theorem:

$$\hat{y} = \underset{y}{\operatorname{argmax}} \left[(1 - \lambda) \log p(y|x) + \lambda \log p(x|y) \right]$$

$p(x|y)$ can be modeled using another seq2seq.

Sequence to sequence: example (MMI, 0.7)

- **Me: Hello**
- **Bot: Hello**
- **Me: What should I do tonight?**
- **Bot: Party**
- **Me: Do you really think so? Where should I go then?**
- **Bot: I don't know. I don't know.**
- **Me: Okay. What should I wear?**
- **Bot: Clothes.**

Me: What do you like more, black jacket or white t-shirt?
Bot: A black shirt.
Me: But I don't have a black shirt!
Bot: I have a black shirt
Me: So can you give it to me, please?
Bot: Sure.

Sequence to sequence: how to generate samples?

So we have our distribution $p(y \mid x)$

How to generate the response y to a question x?

1. Sample from $p(y \mid x)$ with temperature.
2. Find approximate mode with MMI-penalization

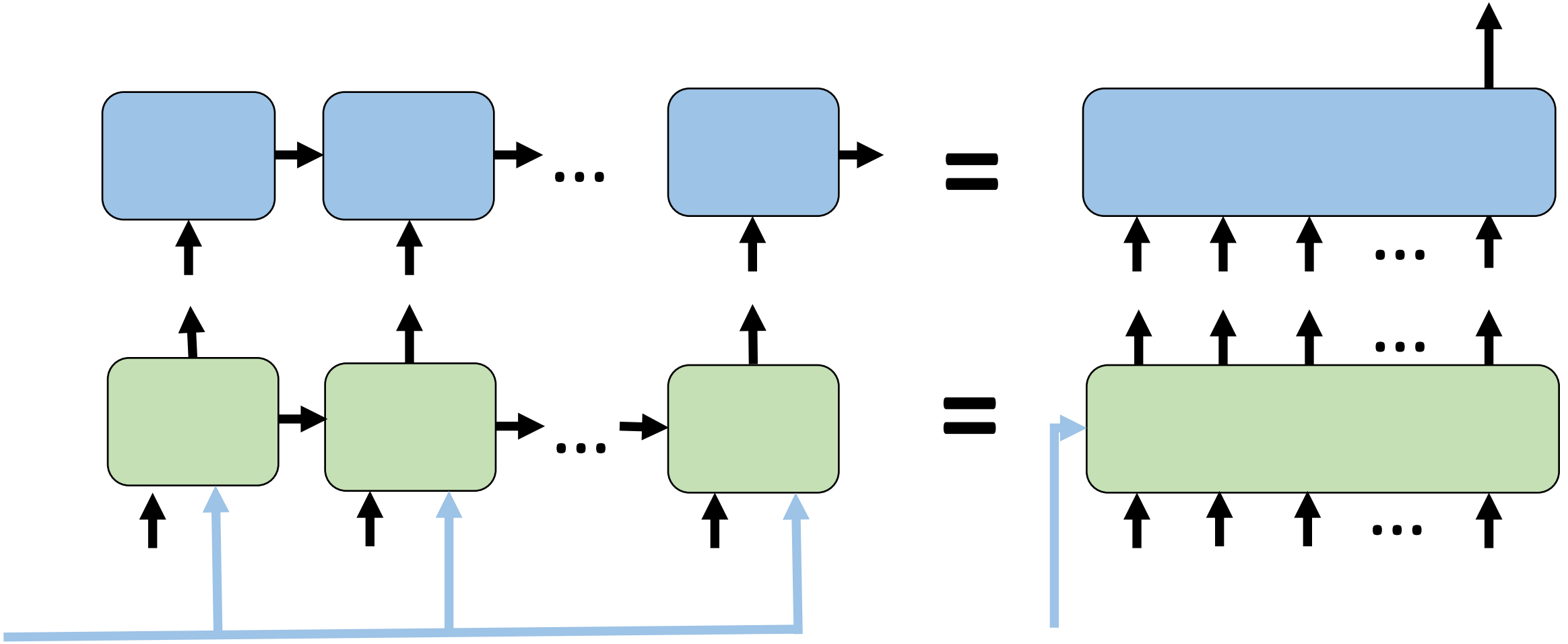
Outline

1. Language Modelling
2. Basic neural conversational models
3. How to generate samples?
4. **Advanced neural conversational models**

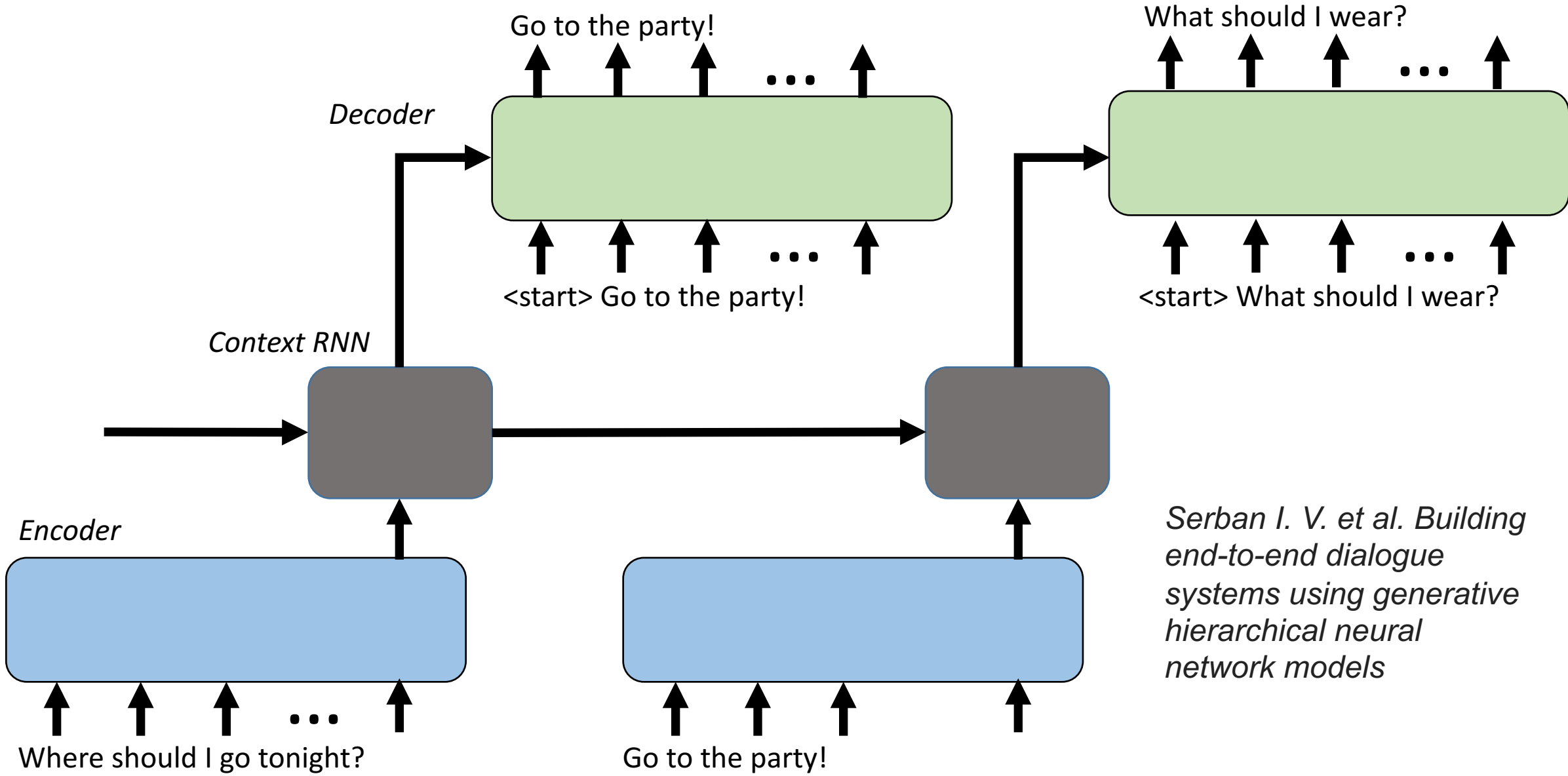
Directions of research

- Hierarchical models
- Models with stochasticity
- Incorporating different goals using RL
- …

HRED: Hierarchical Recurrent Encoder-Decoder

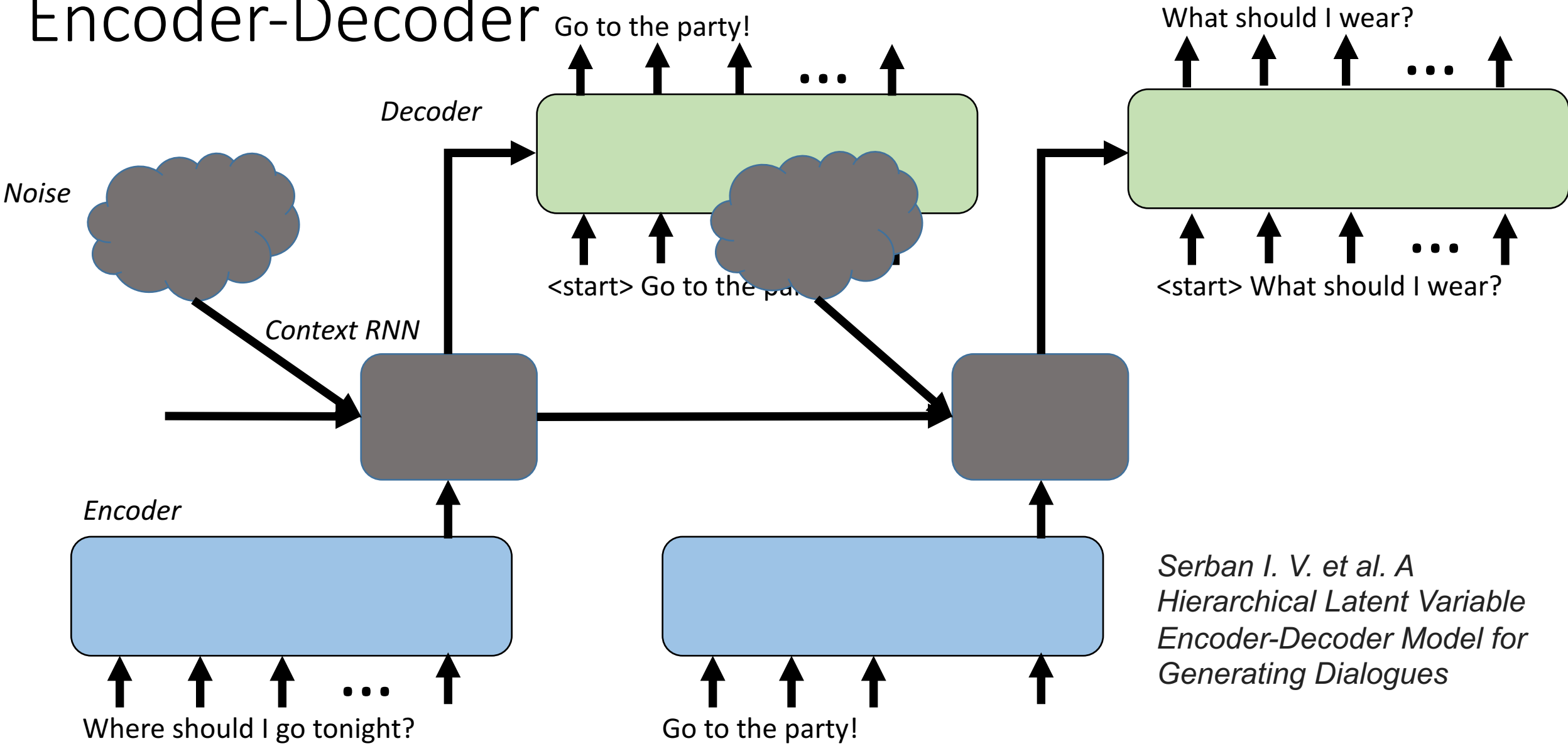


HRED: Hierarchical Recurrent Encoder-Decoder



Serban I. V. et al. Building end-to-end dialogue systems using generative hierarchical neural network models

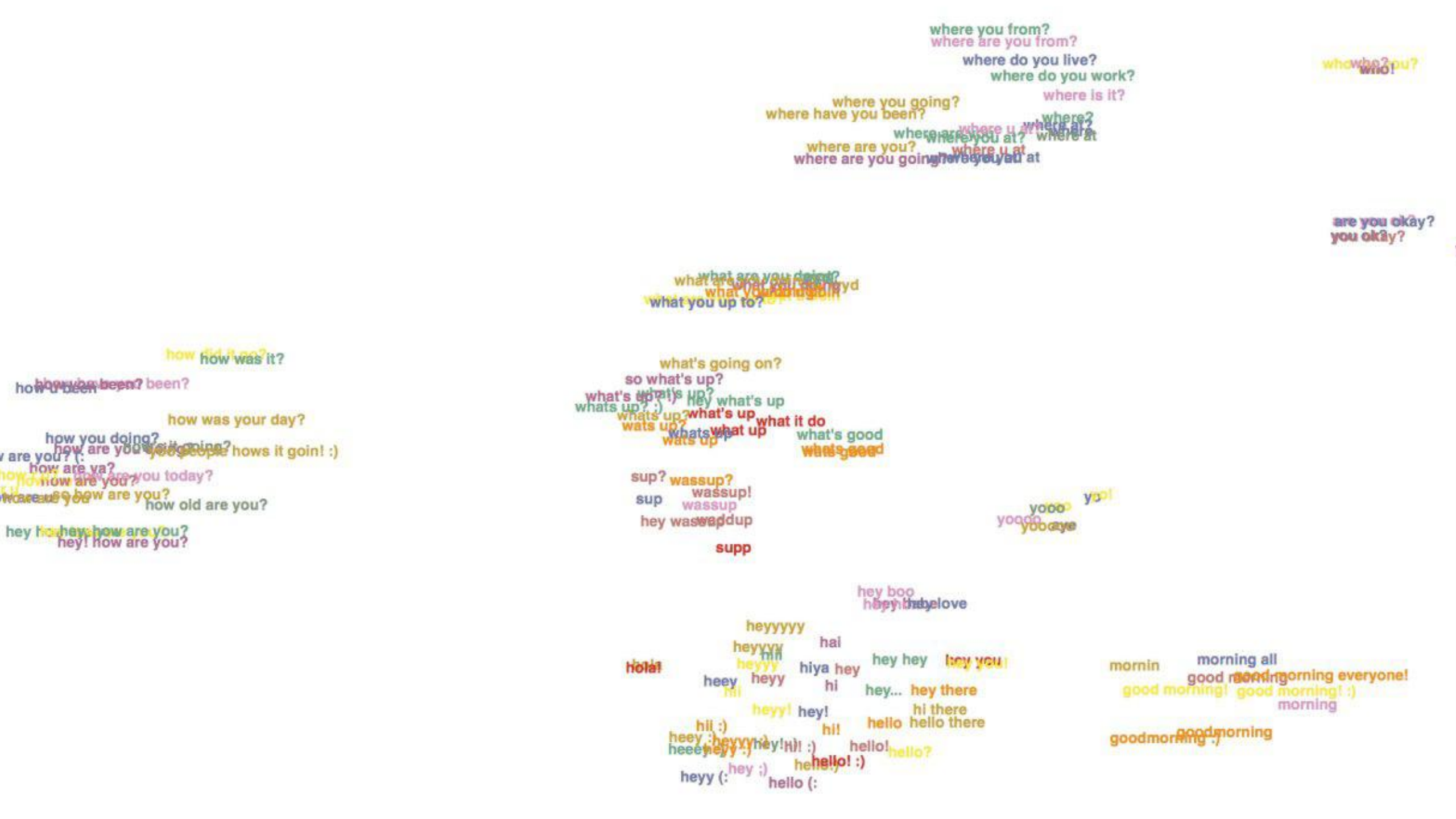
VHRED: Variational Hierarchical Recurrent Encoder-Decoder

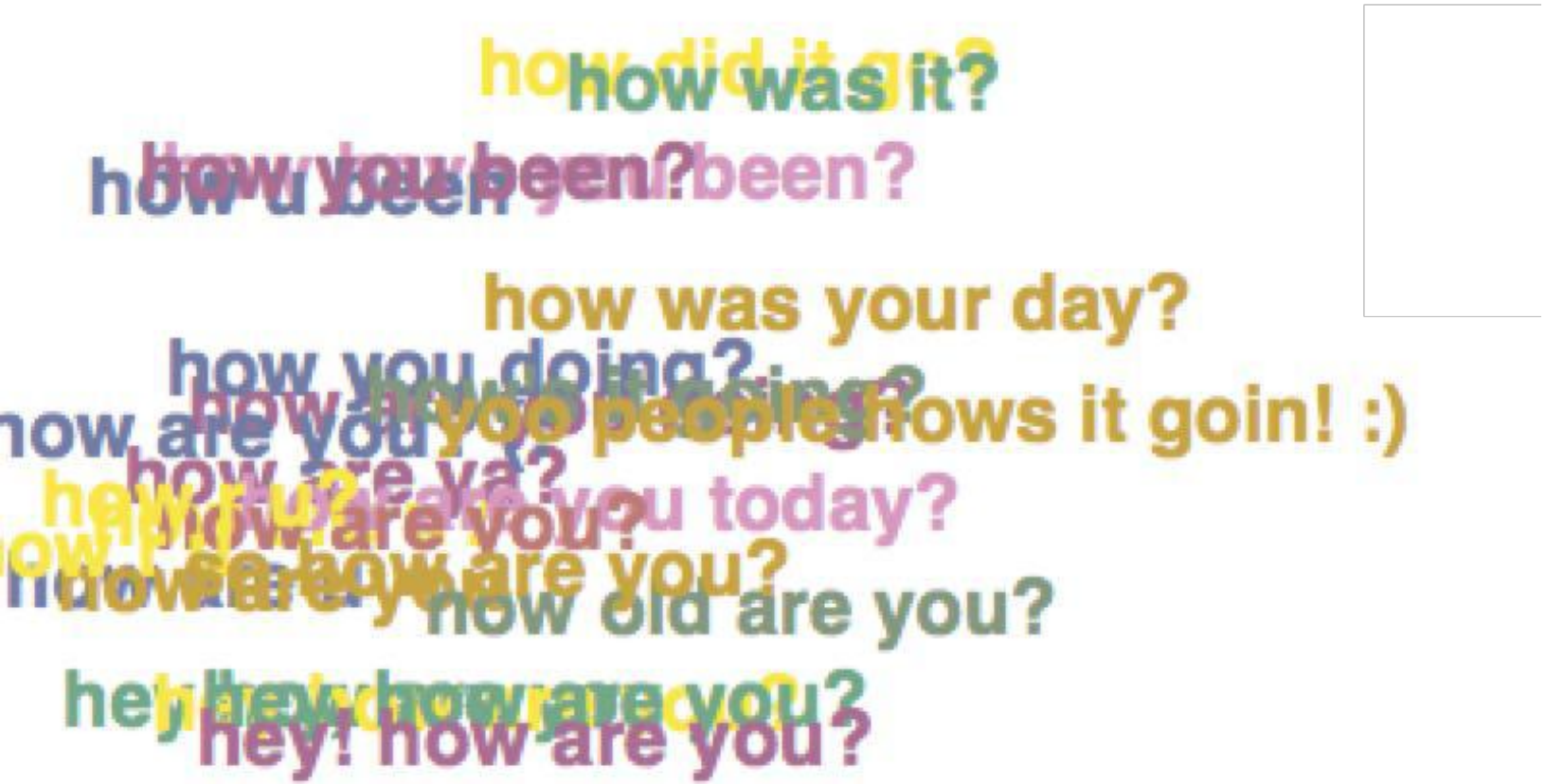


Serban I. V. et al. A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues

Thank you for your attention!


how much?
why's that?
what's wrong?
what happened?
what did you do?
what did he say?
what did he do?
whats that?
lol what
what's it about?
what do you mean?
talking about?
about what?
what time?
what song?
what are you watching?
what do you think?
which one?
what time is it there?
hmm?
where you from?
where are you from?
where do you live?
where do you work?
where is it?
where you going?
where have you been?
who?
when is it?
when!
excuse me?
yes?
hmmm
are you serious?
are you?
what about you?
lies
are you okay?
you okay?
you sure?
good, you?
i'm good and you?
what are you doing
what you up to?
how was it?
how you been?
how was your day?
how you doing?
hows it goin! :)
how are you today?
how old are you?
hey! how are you?
what's going on?
so what's up?
what's up
what it do
what's good
sup
wassup!
hey wassup
supp
hey boo
heyyyyy
hola!
hey there
hi there
hello there
hello!
hello (:
morning all
good morning everyone!
morning
goodmorning
thank you! <3
thanks for the follow :)
thanks for the mention!
thanks!
sorry :(
i'm sorry :(
im bored
never!
not yet
i'm not
i'm back
back
lmao
hahahaha!
hahaha!
lolz
omg
lol shut up
stfu
ew
omg!
lol wow
wow!
that sucks
aww :(
yay!
omg really?
really!
oh ok
lol okay
so?
ok...
bye
kk
fine
ok
hell yea
ikr
yessss
yesss
haha yeah
lol yeah
yepp
right
i know right
yup
exactly!
yeah...
indeed
mhmm
i know
maybe
hmmm
i don't know
yeah it is
ya
this is soooo funny
i hope
idk lol
nope
naw
no :(
yes ma'am
yes i am
i want some
i feel you
just did
good luck
good luck!
you're with me where ever i go
same!
congratulations!
congrats!
you suck
yes you are
you know it
you already know
do you know?
call me
tell me about it
do it
make me
please
me to
me either
me too :)
me too :(
go to sleep
go to bed
go for it!
love it!
looking forward to it!
great
good!
nice!
awesome :)
cool
lucky!
i hate you
i love you
love you!
sure :)
you're welcome
i love you too
i miss you too!
no problem :)
no problem


where you from?
where are you from?
where do you live?
where do you work?
where is it?
where you going?
where have you been?
where?
where at?
where are you?
where are you going?
where at
who?
are you okay?
you okay?
what you up to?
how was it?
how was your day?
how you doing?
how are ya?
how are you?
how old are you?
hey! how are you?
what's going on?
so what's up?
what's up?
hey what's up
what's up
what it do
what's good
sup?
wassup?
wassup!
sup
wassup
supp
yooo
yo!
hey boo
heyyyyy
hai
hola!
hey hey
hey you
heey
heyy
hiya
hey
hi
hey...
hey there
heyy!
hey!
hi there
hii :)
hii
hello
hello there
hello!
hello?
hello! :)
hey ;)
heyy (:
hello (:
mornin
morning all
good morning everyone!
good morning!
good morning! :)
morning
goodmorning
goodmorning :)


how was it?
how you been?
how was your day?
how you doing?
how are you?
how are ya?
how are you today?
how are you?
how old are you?
hey how are you?
hey! how are you?