Attention Based Joint Model with Negative Sampling for New Slot Values Recognition

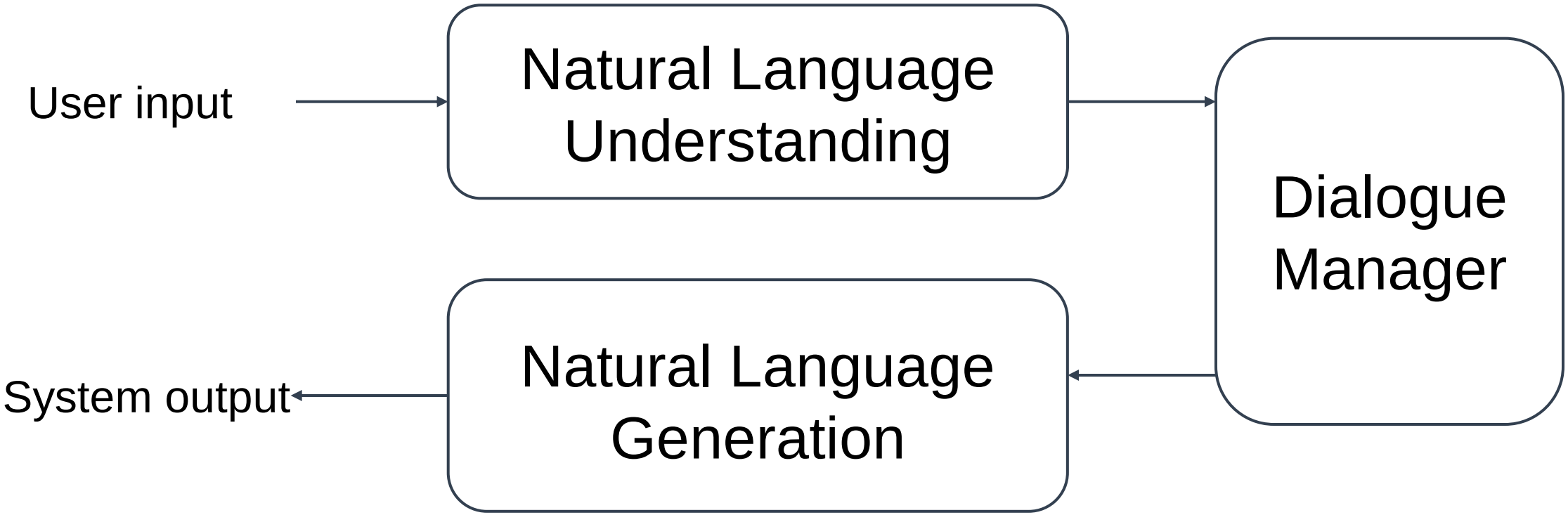
By: Mulan Hou
houmulan@bupt.edu.cn

CONTENTS

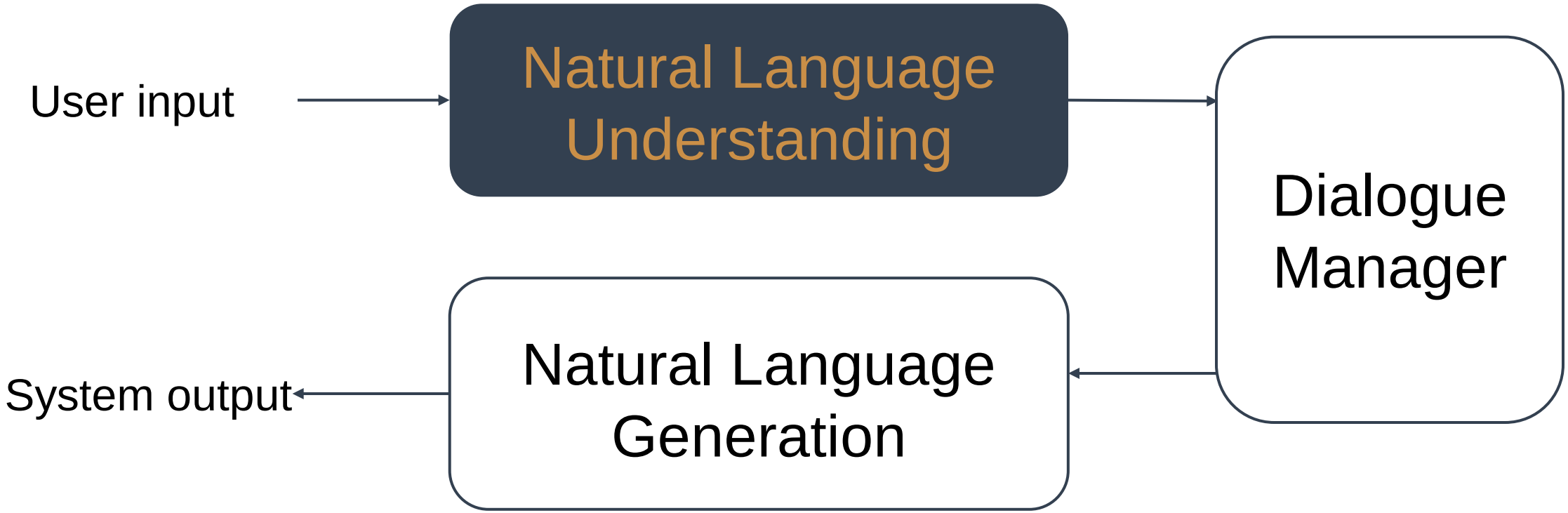
CHAPTER 1

Introduction

Introduction



Introduction

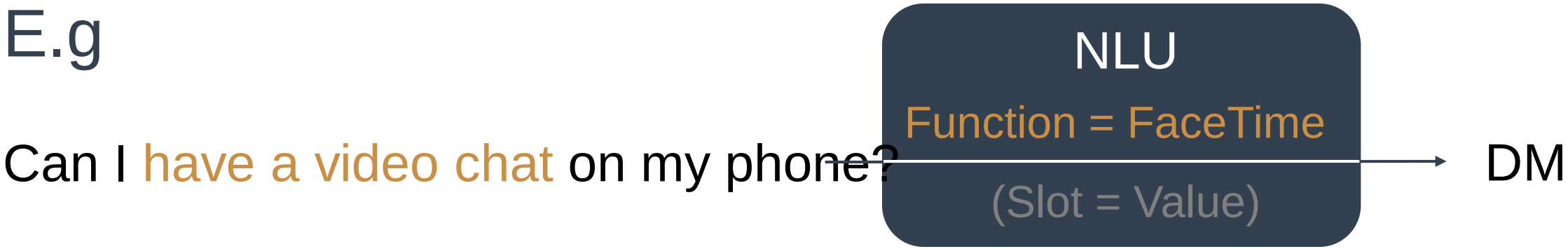


Introduction

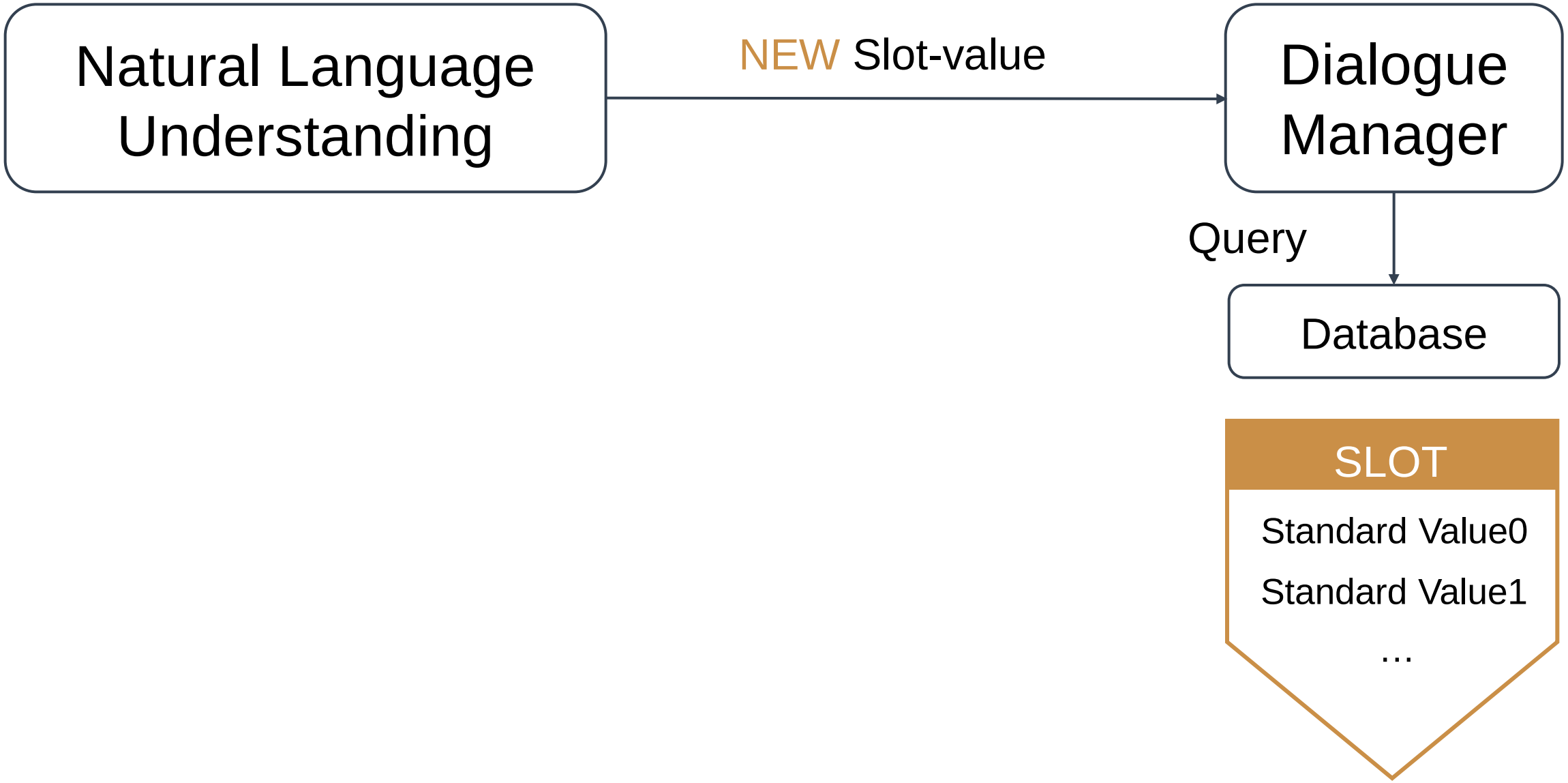
Introduction

Introduction

E.g



Introduction



Introduction

E.g

Can I scroll the screen to have a screenshot on my phone?

Function = Smart Screenshot

(Slot = Value)

Not in the predefined set

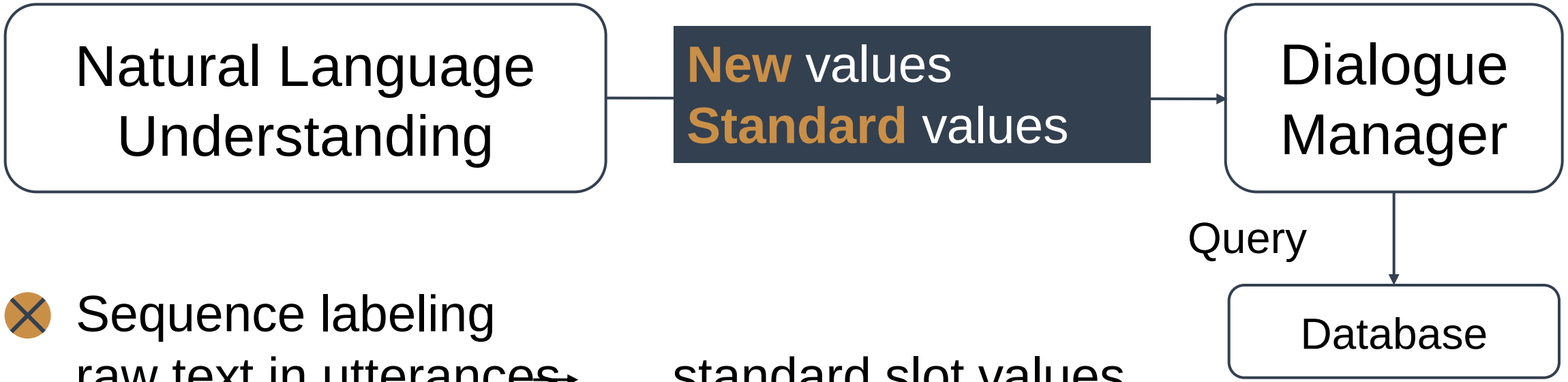
FUNCTION

Face Time

Wechat

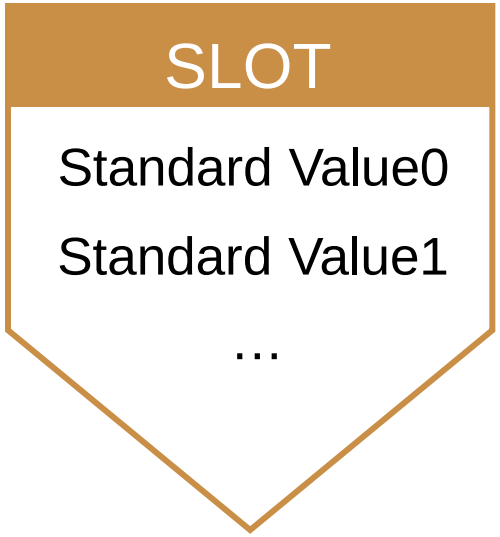
…

Introduction



- Sequence labeling
  raw text in utterances → standard slot values
- Classification

  Problems of new slot values recognition for the lack of training data.
- Attention based joint model with negative sampling



CHAPTER 2

Related work

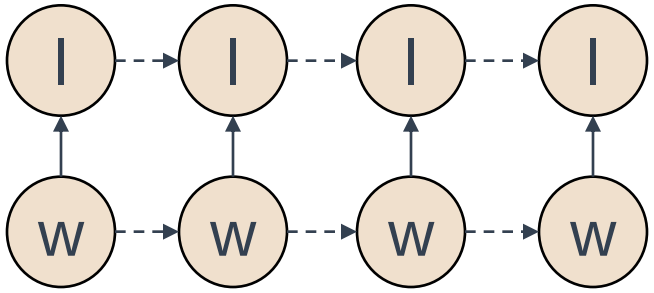
Sequence labeling
Pipeline based
Classification based

Related work

Sequence labeling

Sequence labels

words



O O B I I I O O O

Can I have a video chat on my phone?

Pipeline based methods

Can I have a video chat on my phone? 1 extract the raw texts from utterance

have a video chat → FaceTime 2 normalized the texts into standard slot values

Classification based methods

Can I have a video chat on my phone? → FaceTime

Related work

Sequence labeling

- ✗ Extra normalization operations

Pipeline based methods

- ✗ Prone to accumulating errors

Classification based methods

- ✗
  - Disability to deal with new slot values
  - Losing local information

Related work

Sequence labeling

Xuezhe Ma and Eduard Hovy. *End-to-end sequence labeling via bi-directional lstm-cnns-crf. 2016*

Gökhan Tur, Dilek Hakkani-Tur et al. *Sentence simplification for spoken language understanding. 2011*

Kaisheng Yao, Baolin Peng et al. *Recurrent neural networks for language understanding. 2013*

Kaisheng Yao, Baolin Peng et al. *Spoken language understanding using long short-term memory neural networks. 2014*

Pipeline based methods

F Lef´evre. *Dynamic bayesian networks and discriminative classifiers for multi-stage semantic interpretation. 2007*

Péter Z Yeh, Benjamin Douglas et al. *A speech-driven second screen application for tv program discovery. 2014*

Classification based methods

Rahul Bhagat, Anton Leuski, and Eduard Hovy. *Statistical shallow semantic parsing despite little training data. 2005*

Franc¸ois Mairesse, Milica Gasic et al. *Spoken language understanding from unaligned data using discriminative classification models. 2009*

Ana Mendes Pedro Mota, Lu´ısa Coheur. *Natural language understanding as a classification process: report of initial experiments and results. 2012*

CHAPTER 3

Motivation

Motivation

Sequence labeling classification based methods

- Take local information into consideration
- Obtains normalized slot values directly

Attention based joint model with negative sampling

- Able to deal with new slot values

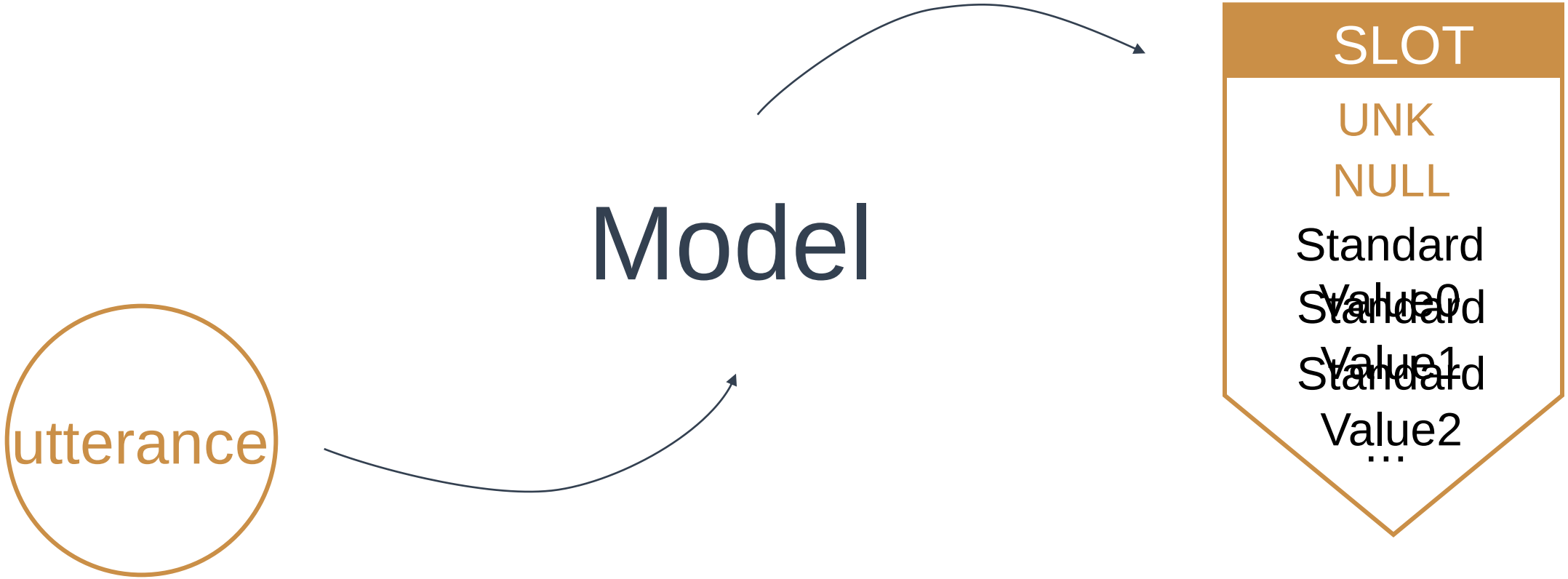
CHAPTE R 4

Proposed model

Attention based joint model
Negative sampling

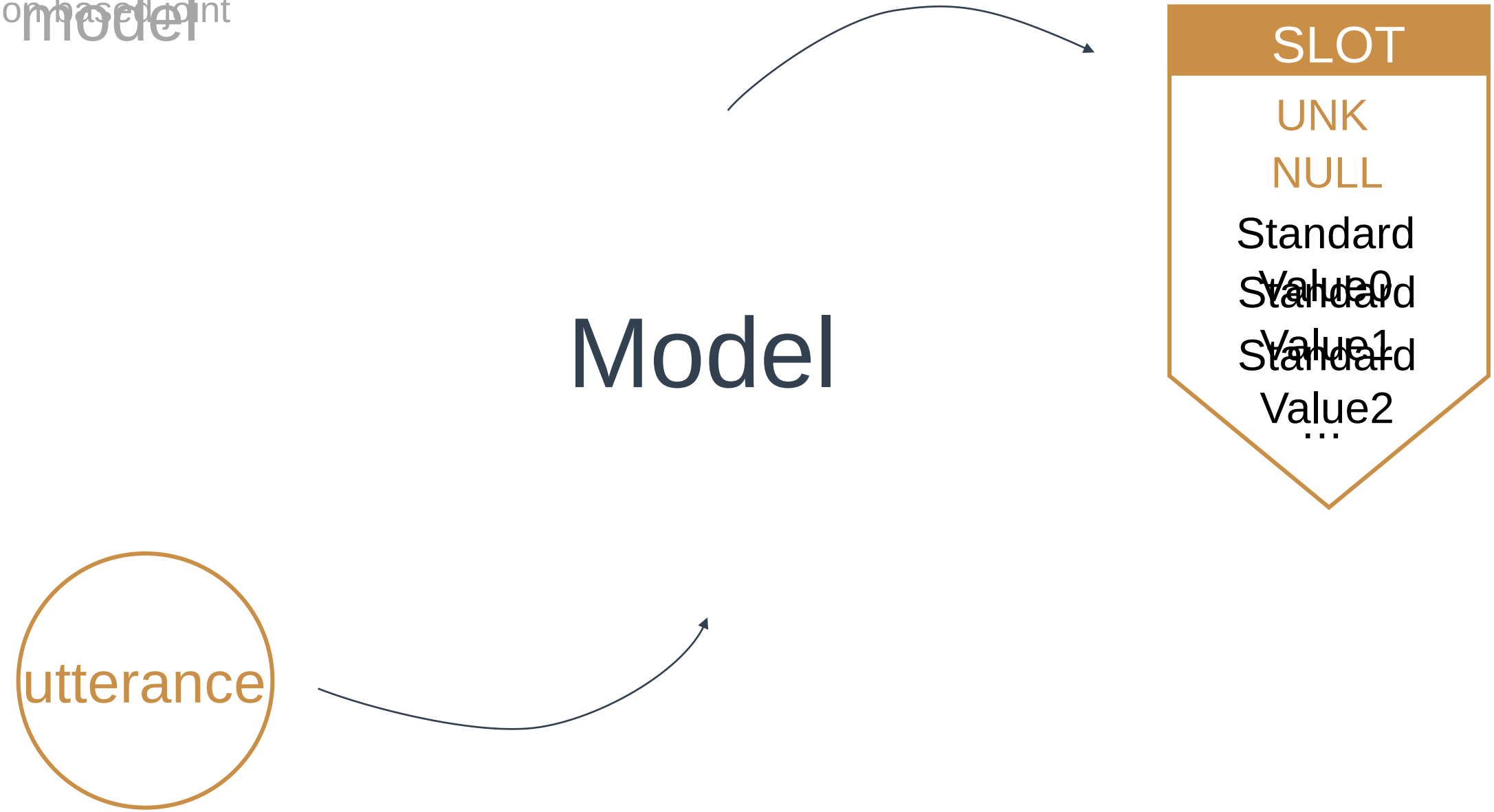
Proposed model

Attention based joint model



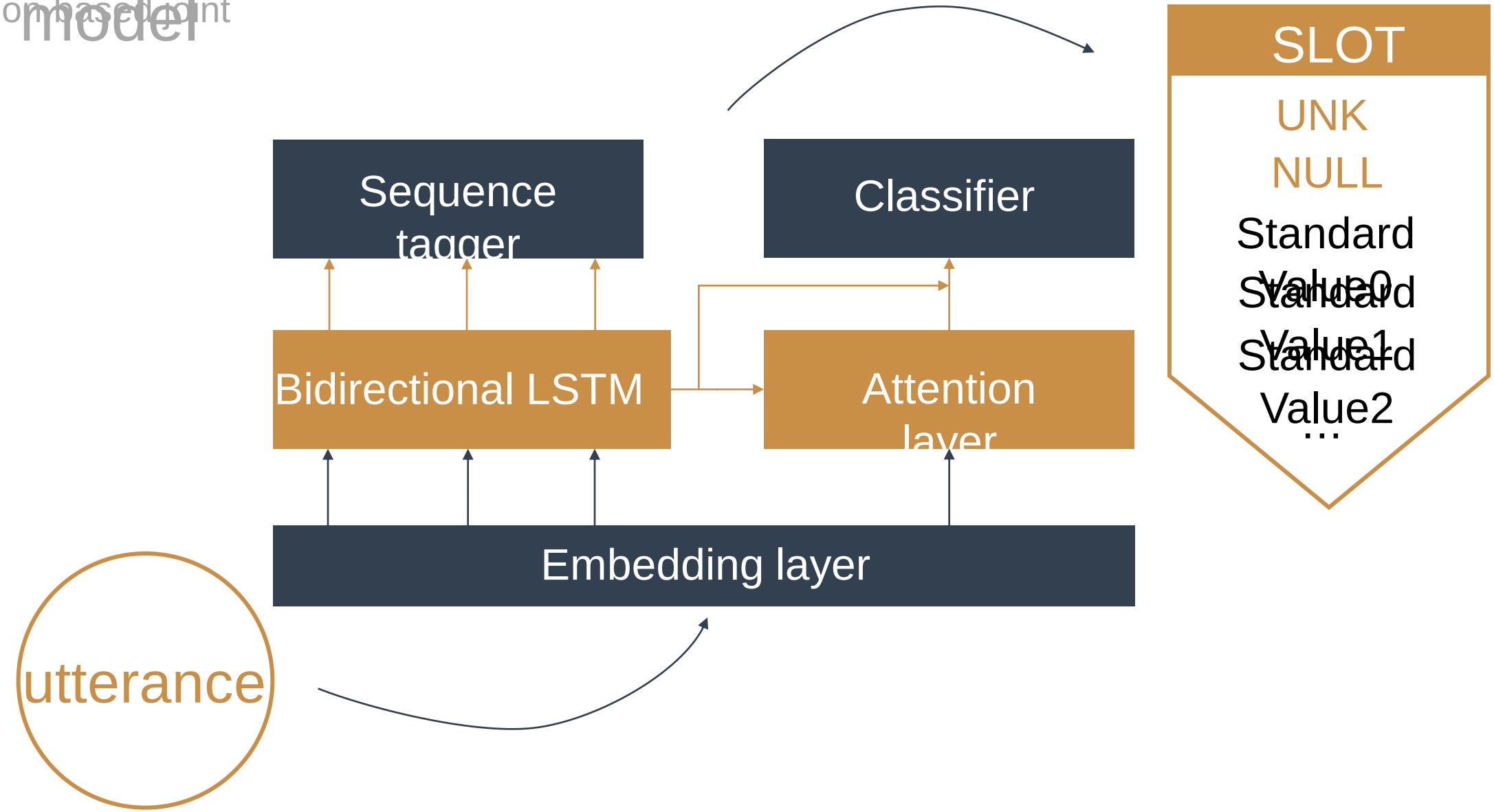
Proposed model

Attention-based joint model



Proposed model

Attention based joint model



Proposed model

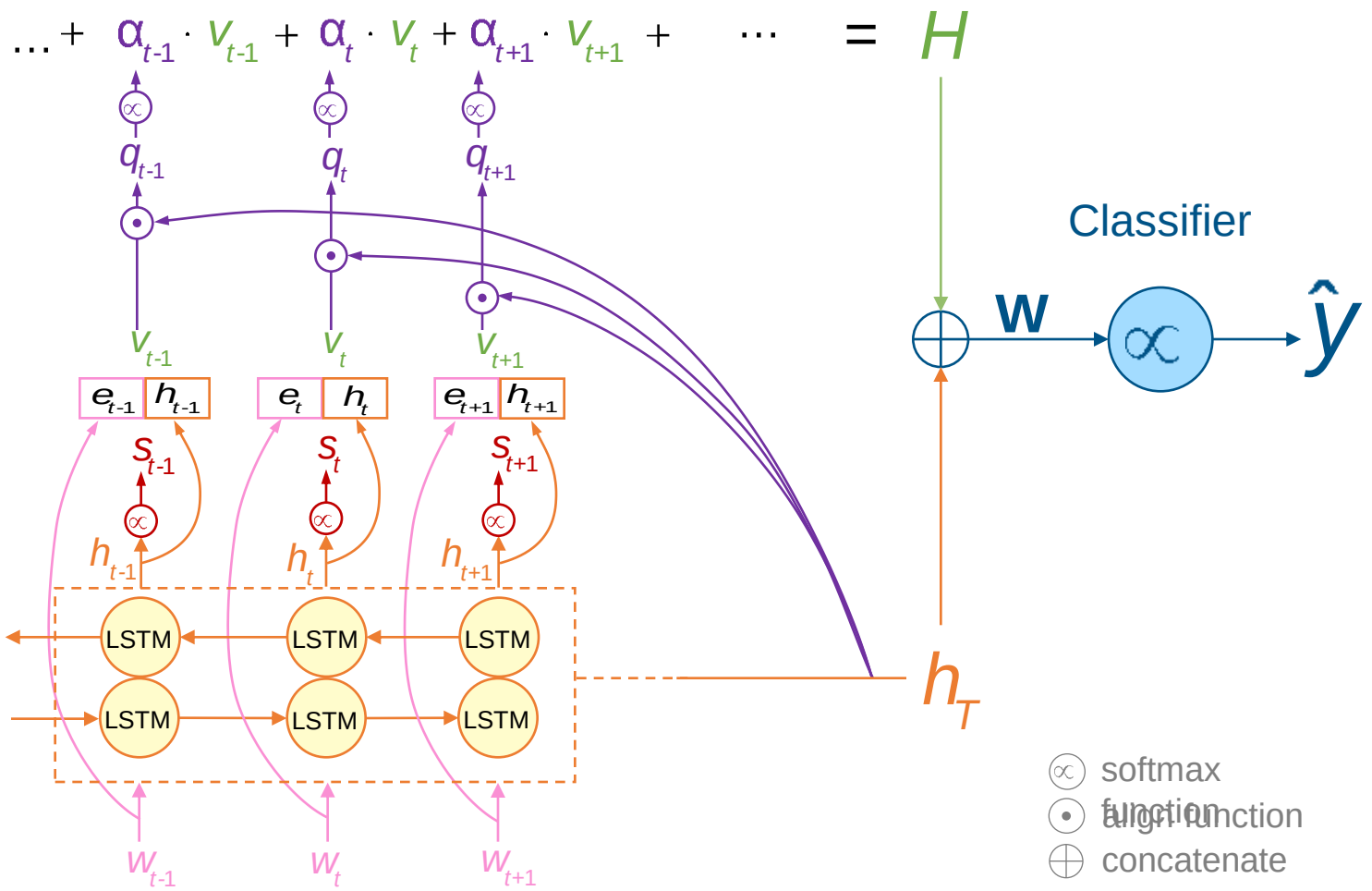
Attention based joint model

Attention layer

Sequence tagger

Bidirectional LSTM

Embedding layer



Proposed model

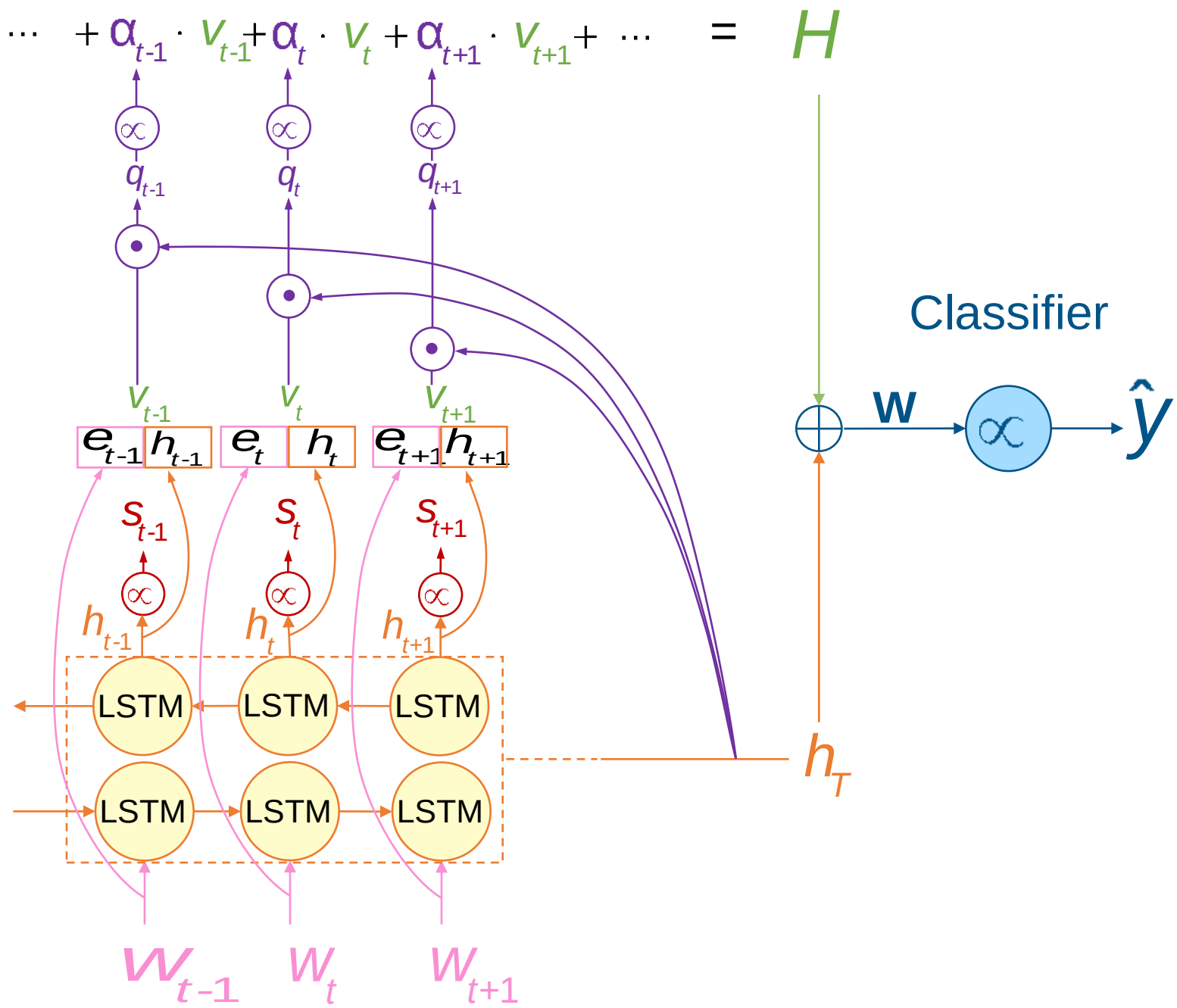
Attention based joint model

Attention layer

Sequence tagger

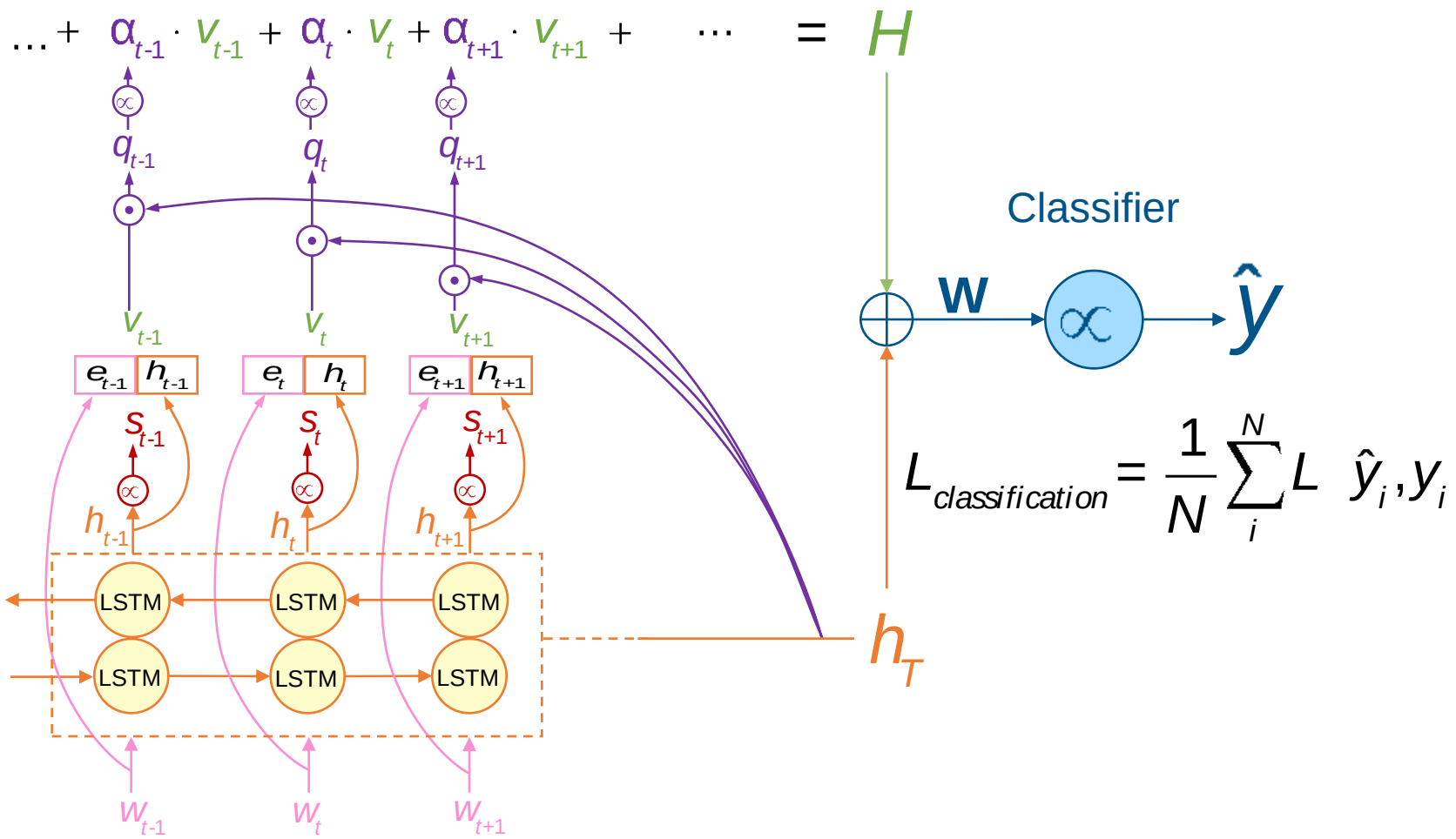
Bidirectional LSTM

Embedding layer



Proposed model

Attention based joint model



Sequence tagger

$$L_{tagging} = \frac{1}{N}\sum_{i}^{N}\frac{1}{T_i}\sum_{t}^{T_i} L\left(\hat{s}_t^i, s_t^i\right)$$

$$L_{classification} = \frac{1}{N}\sum_{i}^{N} L\left(\hat{y}_i, y_i\right)$$

$$L = gL_{tagging} + (1-g)L_{classification}$$

Proposed model

Negative sampling

Models will fail in recognizing new slot values without corresponding training data

Construct negative samples of the existing slot values to simulate new ones

Existing slot value: Can I have a video chat on my phone?

Slot = FaceTime

New slot value: Can I scroll the screen to have a screenshot on my phone?

Slot = Smart Screenshot

Shared context: **Can I … on my phone?**

Can I Random words on my phone?

Non-value

Proposed model

Negative sampling

Models will fail in recognizing new slot values without corresponding training data

Construct negative samples of the existing slot values to simulate new ones

Existing slot value: Can I have a video chat on my phone?

Slot = FaceTime

New slot value: Can I scroll the screen to have a screenshot on my phone?

Slot = Smart Screenshot

Shared context: Can I Random words **Can I … on my phone?** on my phone? → **Negative sampling**

Non-value

Proposed model

Negative sampling

① Words distribution

$$U(word) = \frac{count(word)}{|Data|}$$

② Old values as templates

Can I have a video chat on my phone?

I'd like to find a restaurant serving Chinese food

.....

③ Sample from vocabulary by words distribution to fill the templates

	Can	I	**have**	**a**	**video**	**chat**	on	my	phone	?		FaceTime
	O	O	**B-func**	**I-func**	**I-func**	**I-func**	O	O	O	O		
negative sampling:	**You**	**use**	**battery**	**let**								
	Can	I	**You**	**use**	**battery**	**let**	on	my	phone	?		UNK
	O	O	**B-func**	**I-func**	**I-func**	**I-func**	O	O	O	O		

	我	的	手	机	可	以	**边**	**打**	**电**	**话**	**边**	**视**	**频**	吗	?	视频通话
	O	O	O	O	O	O	**B-func**	**I-func**	**I-func**	**I-func**	**I-func**	**I-func**	**I-func**	O	O	
negative sampling:	**手**	**速**	**录**	**支**	**无**	**用**	**间**									UNK
	我	的	手	机	可	以	**手**	**速**	**录**	**支**	**无**	**用**	**间**	吗	?	
							B-									

CHAPTER 5

Experiments

Results
Analyses

Experiments

Results

Dataset: DSTC(English) | Service(Chinese)

DSTC --- an English dataset from a public contest and we use DSTC2 and DSTC3 together. It collects 5510 dialogues about hotels and restaurants booking. Only the slot *"area"*

Service --- a Chinese dialogue dataset which is mainly about consultation for cell phones and contains a single slot named *'function'*.

Corpus		DSTC			Service		
		train	dev	test	train	dev	test
Original data	old	2805	937	917	3682	514	1063
	new	0	113	275	0	15	64
	null	2244	840	953	427	64	109
negative samples		561	0	0	736	0	0
overall size		5613	1890	2145	4845	593	1236

Statistics of two datasets

Corpus	DSTC			Service		
	train	dev	test	train	dev	test
old	66	64	65	80	55	67
new	0	21	21	0	13	44

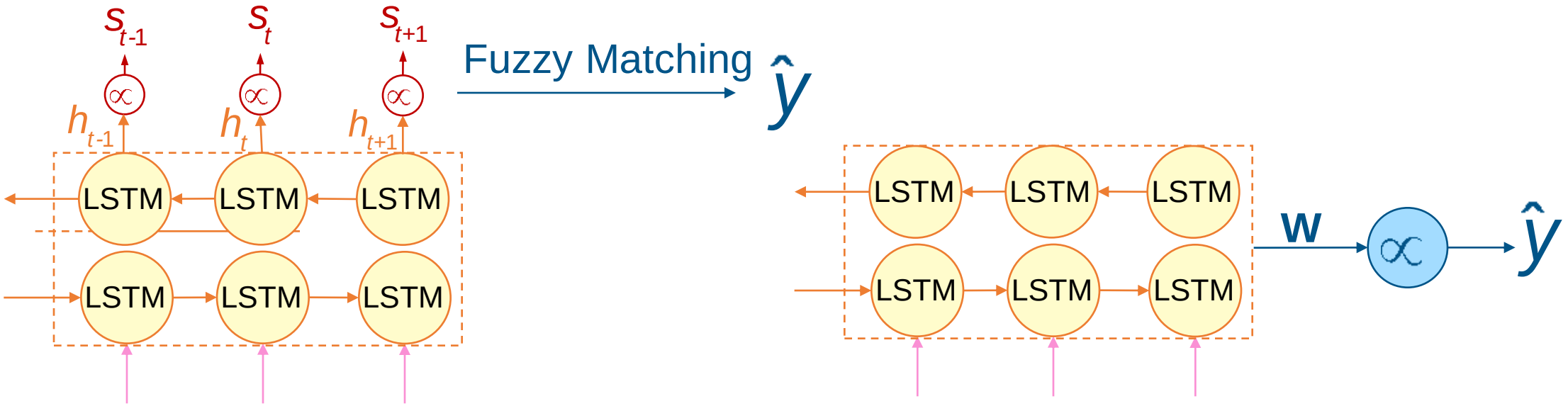
Statistics of value types

Experiments

Results

Baselines: LSTM_FM | LSTM_C, without negative samples

- LSTM_FM: pipeline based method, labeling the words with slot values tags by LSTM model and then normalized them into standard values by Fuzzy Matching.
- LSTM_C: classification based mothed, encoding the utterance by LSTM model and then use a full-connected layer as a Classifier.



LSTM_FM LSTM_C

Experiments

Results

Results:

	DSTC				Service			
	all	NEW	OLD	NULL	all	NEW	OLD	NULL
LSTM_FM	0.8491	0.3063	0.9670	0.8923	0.8981	0.5693	0.9320	0.5693
LSTM_C	0.8290	0.0000	0.9249	0.9761	0.9210	0.0000	0.9720	**0.9643**
AJM_NS(ours)	**0.9632**	**0.8621**	**0.9738**	**0.9822**	**0.9749**	**0.7759**	**0.9881**	0.9633

(a) F1 scores of classification for different models

Experiments

Results

Results:

	DSTC				Service			
	all	NEW	OLD	NULL	all	NEW	OLD	NULL
LSTM_FM	0.8491	0.3063	0.9670	0.8923	0.8981	0.5693	0.9320	0.5693
LSTM_C	0.8290	0.0000	0.9249	0.9761	0.9210	0.0000	0.9720	**0.9643**
	0.9632	**0.8621**	**0.9738**	**0.9822**	**0.9749**	**0.7759**	**0.9881**	0.9633

(a) F1 scores of classification for different models

Experiments

Results

Results:

	DSTC				Service			
	all	NEW	OLD	NULL	all	NEW	OLD	NULL
LSTM_FM	0.8491	0.3063	0.9670	0.8923	0.8981	0.5693	0.9320	0.5693
LSTM_C	0.8290	0.0000	0.9249	0.9761	0.9210	0.0000	0.9720	**0.9643**
AJM_NS(ours)	**0.9632**	**0.8621**	**0.9738**	**0.9822**	**0.9749**	**0.7759**	**0.9881**	0.9633

(a) F1 scores of classification for different models

	DSTC			Service		
	all	NEW	OLD	all	NEW	OLD
LSTM_FM	0.8546	0.2363	0.9837	0.8850	0.2615	0.9269
AJM_NS(ours)	**0.9024**	**0.5684**	**0.9946**	**0.9121**	**0.3399**	**0.9673**

(b) F1 scores of sequence labeling

Experiments

Results

Results:

	DSTC				Service			
	all	NEW	OLD	NULL	all	NEW	OLD	NULL
LSTM_FM	0.8491	0.3063	0.9670	0.8923	0.8981	0.5693	0.9320	0.5693
LSTM_C	0.8290	0.0000	0.9249	0.9761	0.9210	0.0000	0.9720	**0.9643**
AJM_NS(ours)	**0.9632**	**0.8621**	**0.9738**	**0.9822**	**0.9749**	**0.7759**	**0.9881**	0.9633

(a) F1 scores of classification for different models

	DSTC			Service		
	all	NEW	OLD	all	NEW	OLD
LSTM_FM	0.8546	0.2363	0.9837	0.8850	0.2615	0.9269
	0.9024	**0.5684**	**0.9946**	**0.9122**	**0.3399**	**0.9673**

(b) F1 scores of sequence labeling

Experiments

Results

Comparison inside model: Attention mechanism(AJM) & Negative sampling(JM_NS)

	DSTC				Service			
	all	NEW	OLD	NULL	All	NEW	OLD	NULL
Full(AJM_NS)	**0.9632**	**0.8621**	0.9738	**0.9822**	**0.9749**	**0.7759**	**0.9881**	**0.9633**
-Attention only(JM_NS)	0.9515	0.8129	**0.9739**	0.9699	0.9700	0.7207	0.9862	0.9585
-NS only(AJM)	0.8247	0.0000	0.9426	0.9492	0.9234	0.0000	0.9761	0.9511

Experiments

Results

Comparison inside model: Attention mechanism(AJM) & Negative sampling(JM_NS)

	DSTC				Service			
	all	NEW	OLD	NULL	All	NEW	OLD	NULL
Full(AJM_NS)	**0.9632**	**0.8621**	**0.9738**	**0.9822**	**0.9749**	**0.7759**	**0.9881**	**0.9633**
-NS only(AJM)	0.8247	0.0000	0.9426	0.9492	0.9234	0.0000	0.9761	0.9511

- Confusion metrices

AJM → AJM_NS

DSTC

	NEW	OLD	NULL
NEW	0	184	91
OLD	0	916	1
NULL	0	9	944

➡

	NEW	OLD	NULL
NEW	225	33	17
OLD	9	908	0
NULL	13	4	936

Service

	NEW	OLD	NULL
NEW	0	55	9
OLD	0	1063	0
NULL	0	2	107

➡

	NEW	OLD	NULL
NEW	45	17	2
OLD	4	1057	2
NULL	3	1	105

Experiments

Results

Comparison inside model: Attention mechanism(AJM) & Negative sampling(JM_NS)

	DSTC				Service			
	all	NEW	OLD	NULL	All	NEW	OLD	NULL
Full(AJM_NS)	**0.9632**	**0.8621**	**0.9738**	**0.9822**	**0.9749**	**0.7759**	**0.9881**	**0.9633**
-NS only(AJM)	0.8247	0.0000	0.9426	0.9492	0.9234	0.0000	0.9761	0.9511

- Classification results based on negative samples

	DSTC				Service			
	all	NEW	OLD	NULL	all	NEW	OLD	NULL
LSTM_FM_NS	0.8572	0.3536	0.9286	0.9241	0.9642	0.6203	0.9009	0.6488
LSTM_C_NS	0.9543	0.8261	0.9637	**0.9822**	0.9684	0.7103	0.9825	**0.9815**
JM_NS	0.9515	0.8129	**0.9739**	0.9699	0.9700	0.7207	0.9862	0.9585

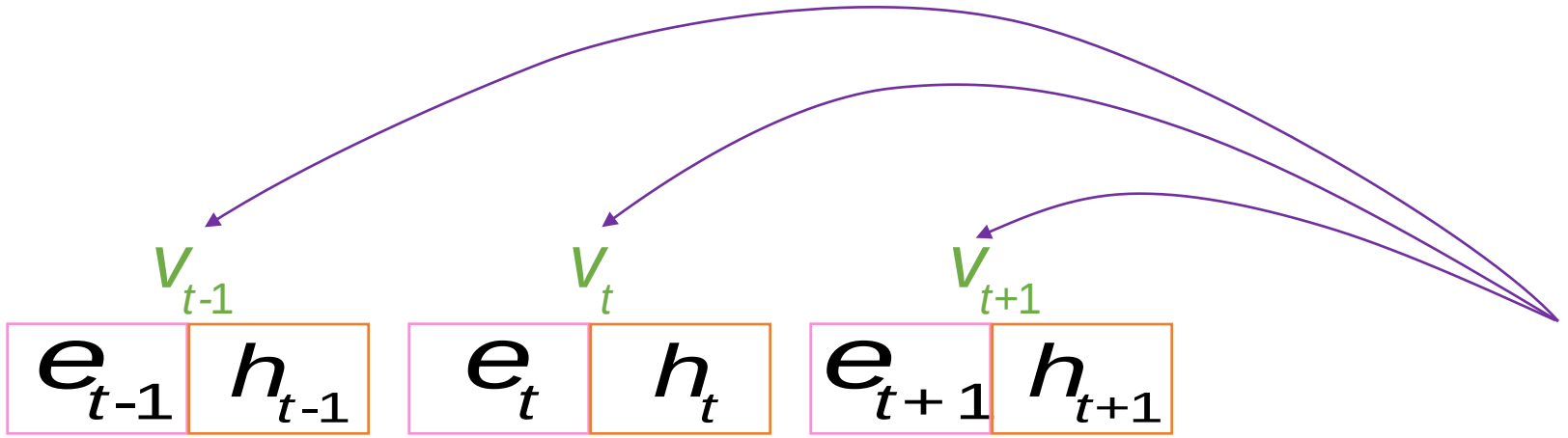
Experiments

Results

Comparison inside model: Attention mechanism(AJM) & Negative sampling(JM_NS)

	DSTC				Service			
	all	NEW	OLD	NULL	All	NEW	OLD	NULL
Full(AJM_NS)	**0.9632**	**0.8621**	0.9738	**0.9822**	**0.9749**	**0.7759**	**0.9881**	**0.9633**
-Attention only(JM_NS)	0.9515	0.8129	**0.9739**	0.9699	0.9700	0.7207	0.9862	0.9585

- Attention mechanism



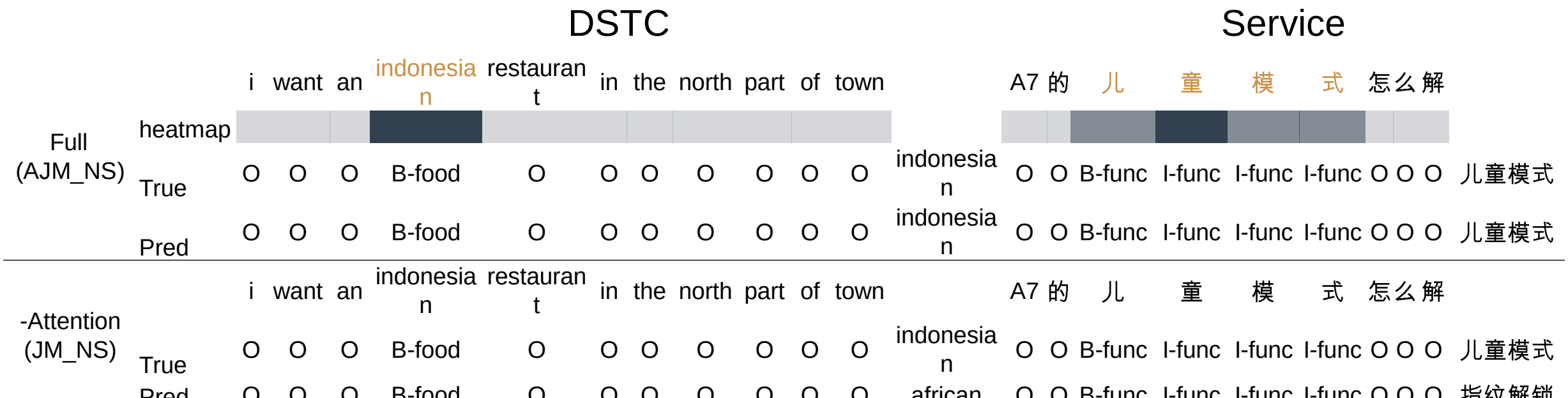
Experiments

Results

Comparison inside model: Attention mechanism(AJM) & Negative sampling(JM_NS)

	DSTC				Service			
	all	NEW	OLD	NULL	All	NEW	OLD	NULL
Full(AJM_NS)	**0.9632**	**0.8621**	0.9738	**0.9822**	**0.9749**	**0.7759**	**0.9881**	**0.9633**
-Attention only(JM_NS)	0.9515	0.8129	**0.9739**	0.9699	0.9700	0.7207	0.9862	0.9585

- Attention mechanism



CHAPTER 6

Conclusion

Conclusion

We propose an attention based joint model with negative sampling.

- Maps the utterance into standard slot values directly without extra normalization operations
- Negative sampling for existing values for a certain slot S enables our model to effectively recognize new slot values
- Joint model collaborated by attention mechanism promotes the performance
- Experimental results demonstrate that our model achieves impressive improvements on new slot values with less damage on other sub-datasets

THANK YOU