UNIVERSITY OF CAMBRIDGE

Scalable Neural Language Generation for Spoken Dialogue Systems

XRCE Seminar, 23/02/2016

Tsung-Hsien (Shawn) Wen and Steve Young

Dialogue Systems Group

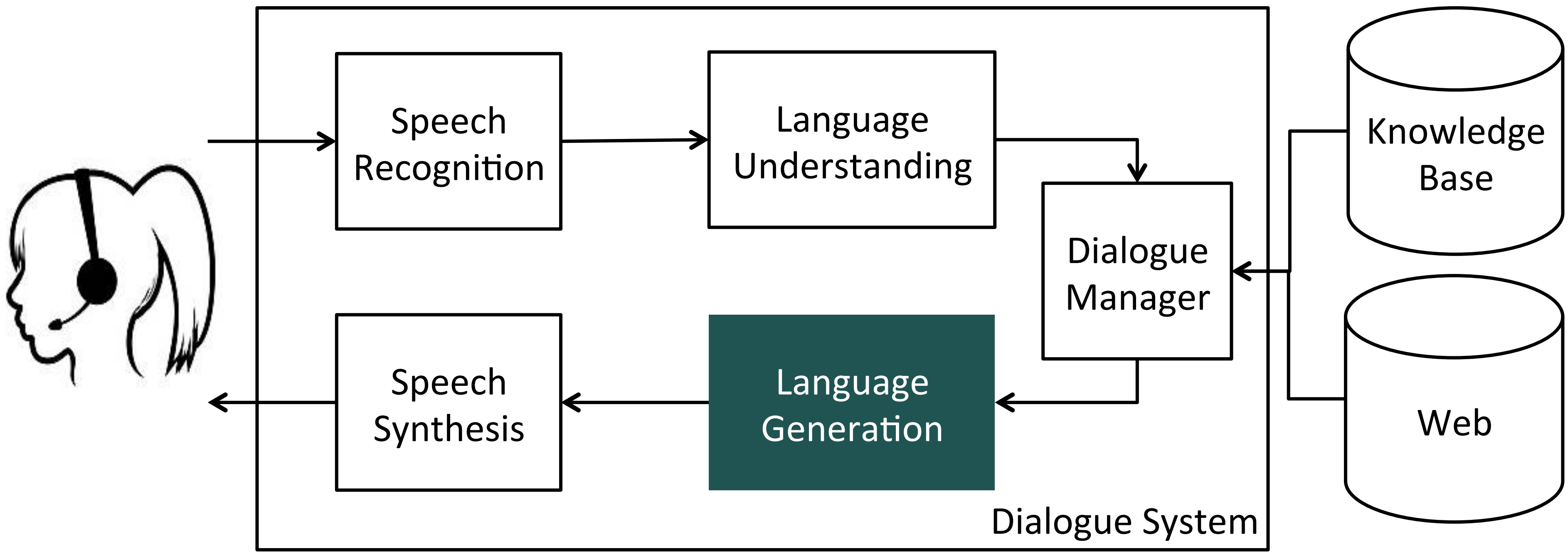
Outline

- Intro
- Semantically Conditioned LSTM
- Domain adaptation for NLG
- Conclusion

Outline

- **Intro**
- Semantically Conditioned LSTM
- Domain adaptation for NLG
- Conclusion

Spoken Dialogue System



NLG: Problem Definition

- Given a meaning representation, map it into natural language utterances.

Dialogue Act

Realisations

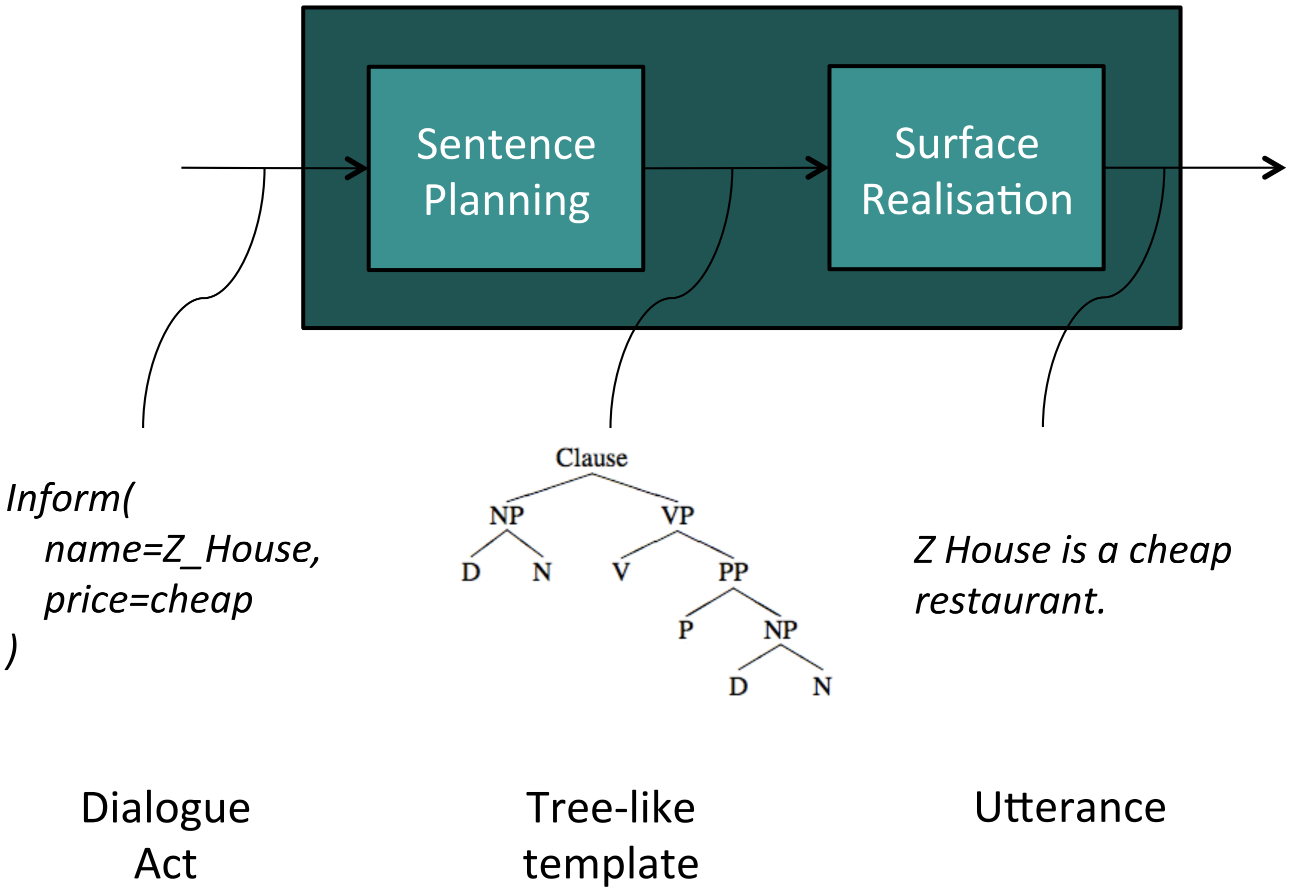
Inform(restaurant=Seven_days, food=Chinese)

Seven days is a restaurant serving Chinese.

Seven days is a Chinese restaurant.

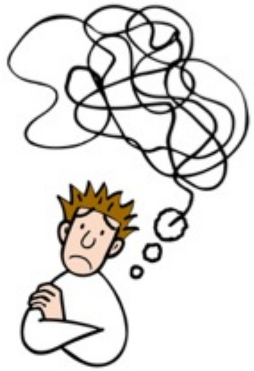
- What do we care about?
  - adequacy, fluency, readability, variation
    (Stent et al 2005)

Traditional pipeline approach



Problems

- Scalability
  - Grammars are handcrafted.
  - Require expert knowledge.



$$
\begin{array}{llllllll}
A \rightarrow & mm, & \Pr(0.11) & \mid\ mh, & \Pr(0.67) & \mid\ hh, & \Pr(0.22) \\
B \rightarrow & mm, & \Pr(0.68) & \mid\ hm, & \Pr(0.23) & \mid\ hh, & \Pr(0.09) \\
C \rightarrow & mm, & \Pr(0.58) & \mid\ hm, & \Pr(0.42) & & \\
\\
T \rightarrow & hQA, & \Pr(0.12) & \mid\ hQB, & \Pr(0.18) & \mid\ APm, & \Pr(0.16) \\
\\
 & & & \mid\ BPh, & \Pr(0.39) & \mid\ hOm, & \Pr(0.15) \\
U \rightarrow & ARC, & \Pr(0.13) & \mid\ BRB, & \Pr(0.44) & \mid\ BRC, & \Pr(0.36) \\
 & & & \mid\ CRC, & \Pr(0.07) & & \\
V \rightarrow & ARA, & \Pr(0.16) & \mid\ ARB, & \Pr(0.66) & \mid\ CRB, & \Pr(0.08) \\
 & & & \mid\ hQA, & \Pr(0.10) & & \\
W \rightarrow & BRA, & \Pr(0.10) & \mid\ CRA, & \Pr(0.08) & \mid\ CRB, & \Pr(0.07) \\
 & & & \mid\ X]\{X, & \Pr(0.75) & & \\
\\
R \rightarrow & lWm, & \Pr(0.14) & \mid\ mWm, & \Pr(0.22) & \mid\ mWh, & \Pr(0.23) \\
 & & & \mid\ hWm, & \Pr(0.17) & \mid\ hWh, & \Pr(0.24)\,' \\
Q \rightarrow & AVh, & \Pr(0.28) & \mid\ BVm, & \Pr(0.55) & \mid\ BVh, & \Pr(0.06) \\
 & & & \mid\ CVh, & \Pr(0.10) & & \\
P \rightarrow & lUB, & \Pr(0.14) & \mid\ mUC, & \Pr(0.22) & \mid\ hUA, & \Pr(0.20) \\
 & & & \mid\ hUC, & \Pr(0.44) & & \\
O \rightarrow & ATA, & \Pr(0.86) & \mid\ CTC, & \Pr(0.14) & & \\
\\
X \rightarrow & xX, & \Pr(0.35) & \mid\ \epsilon, & \Pr(0.65) & & \\
S \rightarrow & [XTX\}, & \Pr(1.00) & & & &
\end{array}
$$

Problems

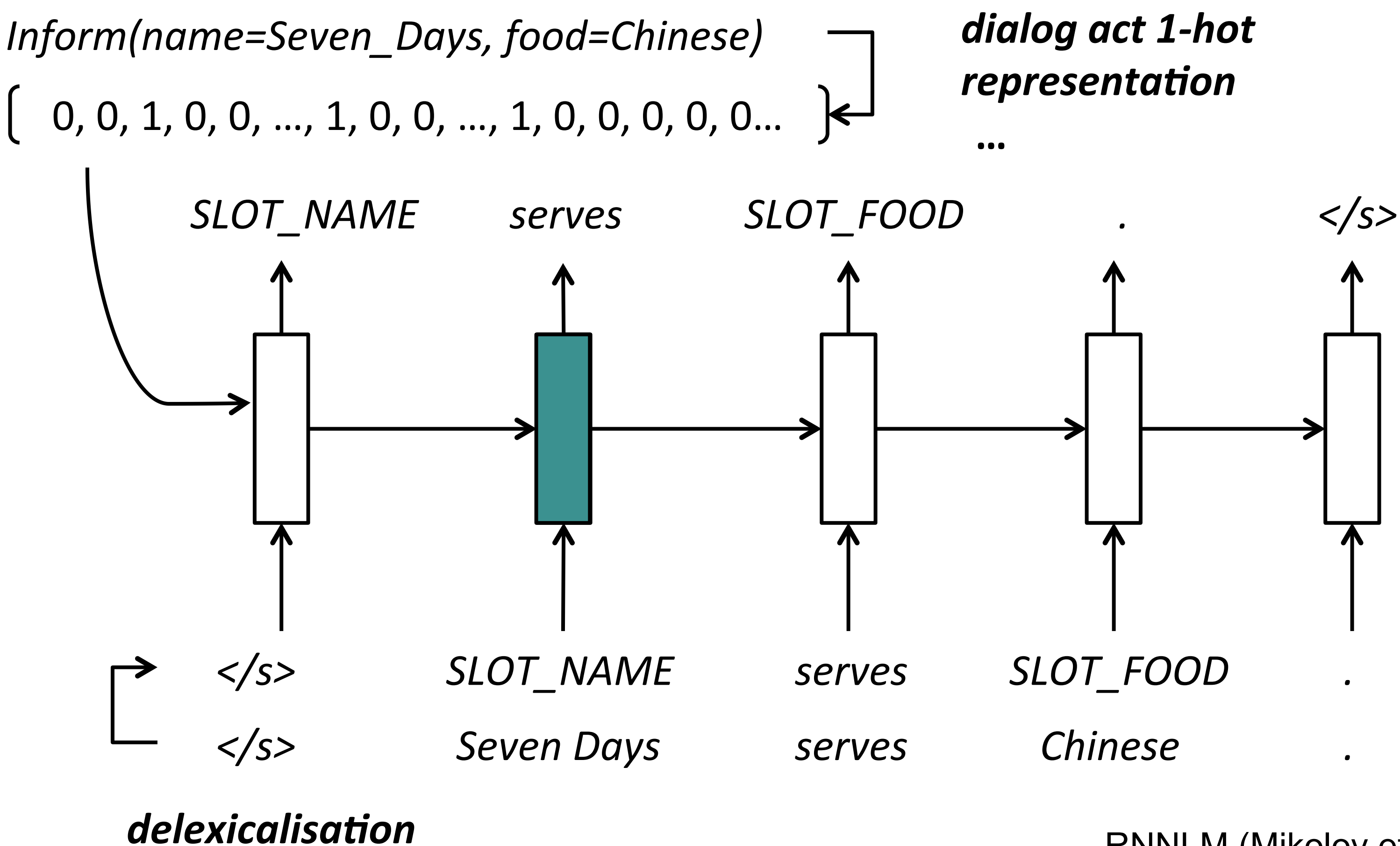
- Boring
  - Frequent repetition of outputs.
  - Non-colloquial, awkward utterances.

Seven Days is a nice restaurant in the expensive price range, in the north part of the town, if you don't care about what food they serve.

Outline

- Intro
- **<u>Semantically Conditioned LSTM</u>**
- Experiments
- Adaptation – A preliminary work
- Conclusion

Recurrent Generation Model



RNNLM (Mikolov et al, 2010)

SC-LSTM

- Original LSTM cell

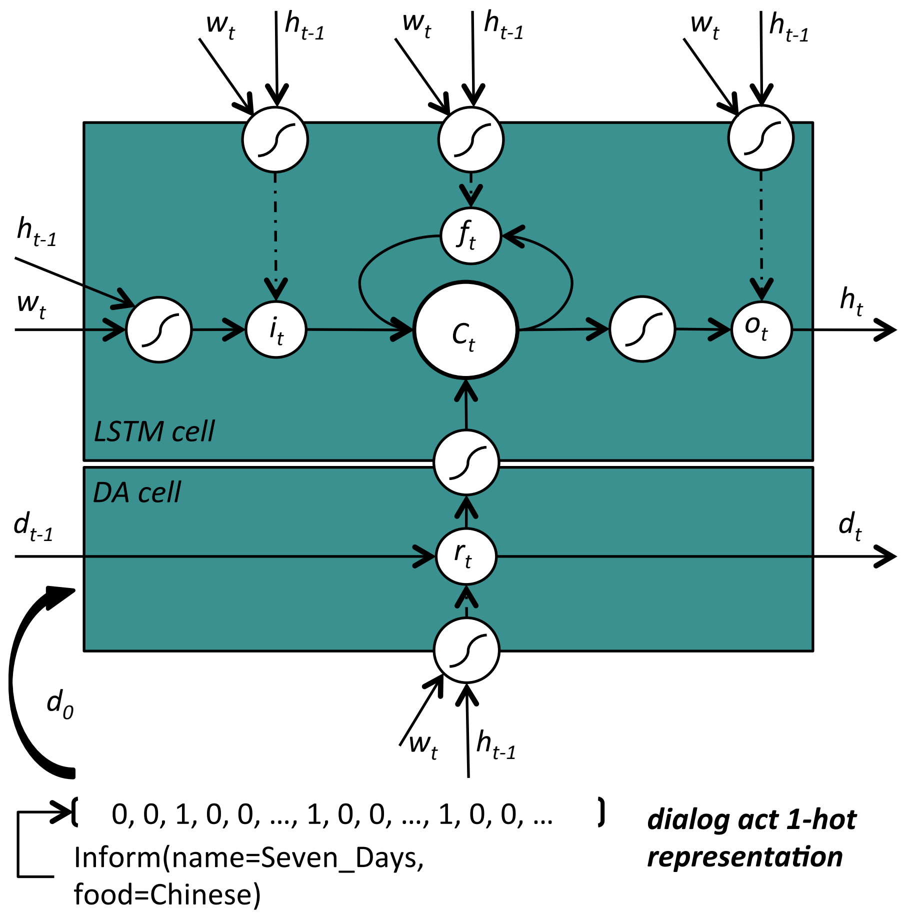
$$\mathbf{i}_t = \sigma(\mathbf{W}_{wi}\mathbf{w}_t + \mathbf{W}_{hi}\mathbf{h}_{t-1})$$
$$\mathbf{f}_t = \sigma(\mathbf{W}_{wf}\mathbf{w}_t + \mathbf{W}_{hf}\mathbf{h}_{t-1})$$
$$\mathbf{o}_t = \sigma(\mathbf{W}_{wo}\mathbf{w}_t + \mathbf{W}_{ho}\mathbf{h}_{t-1})$$
$$\hat{\mathbf{c}}_t = \tanh(\mathbf{W}_{wc}\mathbf{w}_t + \mathbf{W}_{hc}\mathbf{h}_{t-1})$$
$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \hat{\mathbf{c}}_t$$
$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t)$$

- DA cell

$$\mathbf{r}_t = \sigma(\mathbf{W}_{wr}\mathbf{w}_t + \mathbf{W}_{hr}\mathbf{h}_{t-1})$$
$$\mathbf{d}_t = \mathbf{r}_t \odot \mathbf{d}_{t-1}$$

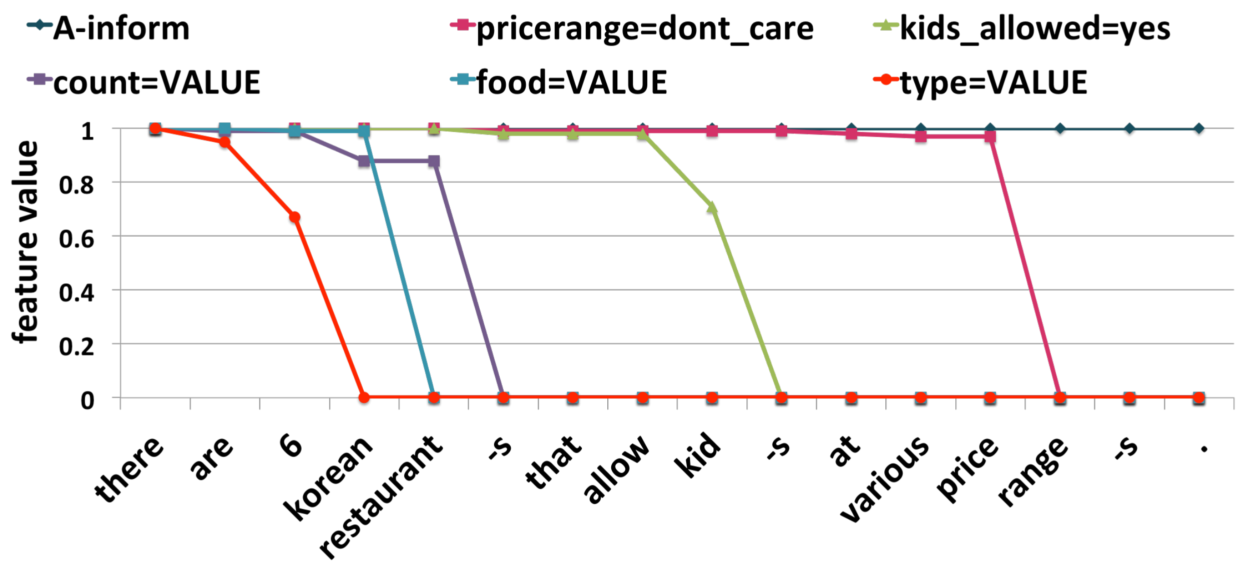
- Modify Ct

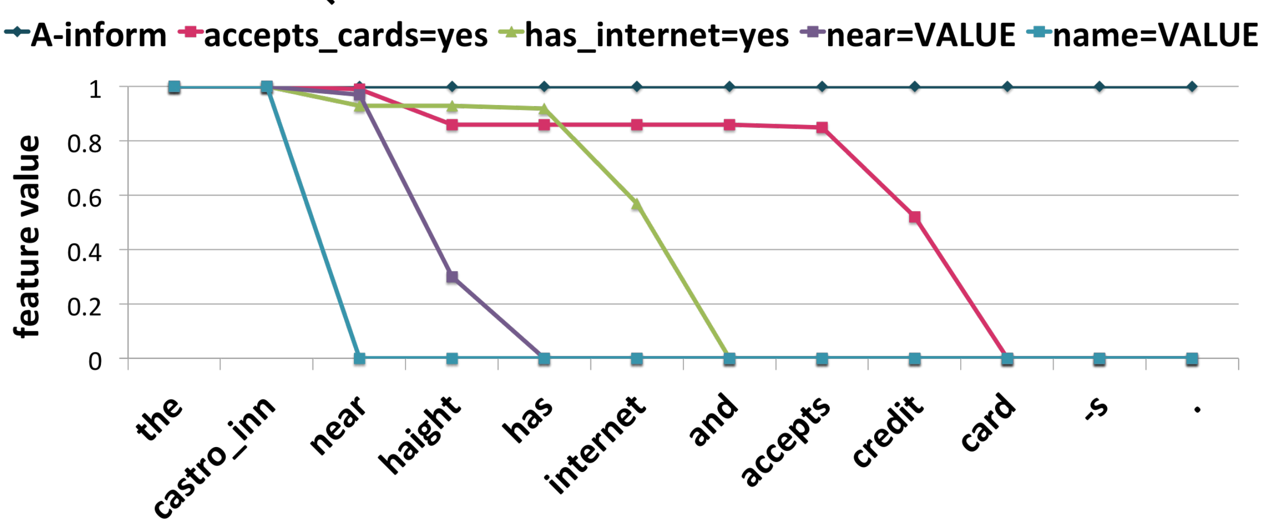
$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \hat{\mathbf{c}}_t + \tanh(\mathbf{W}_{dc}\mathbf{d}_t)$$



(Hochreiter and Schmidhuber, 1997)

Visualization



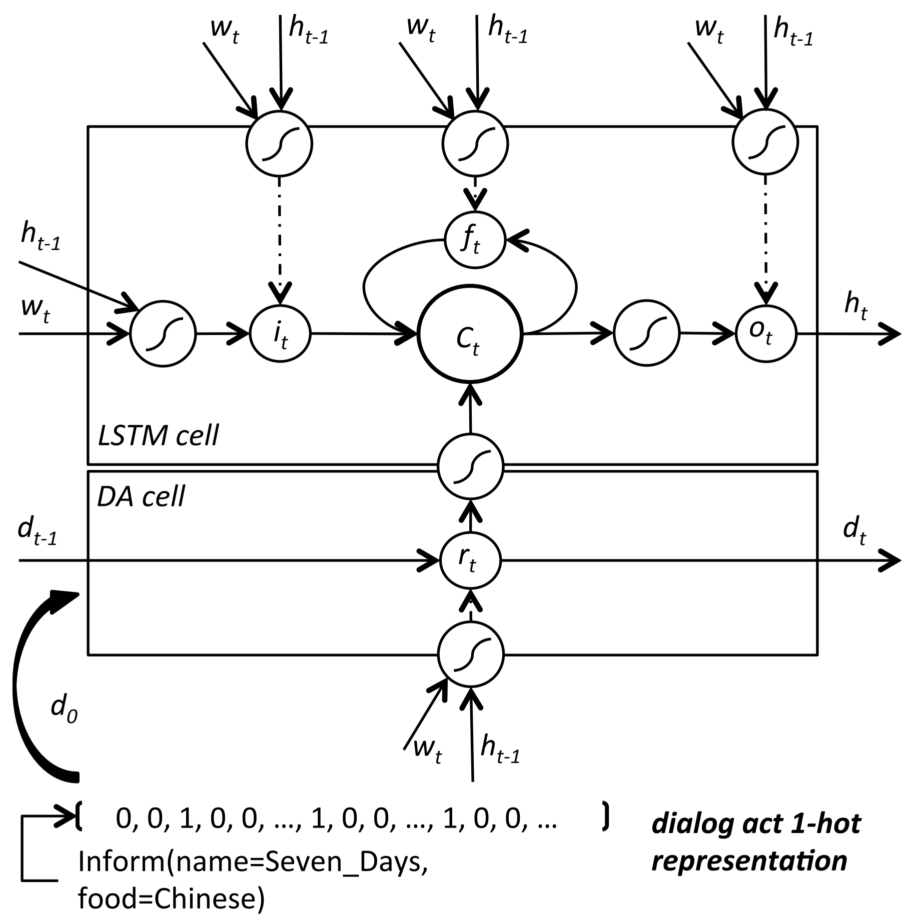


SC-LSTM

- Cost function

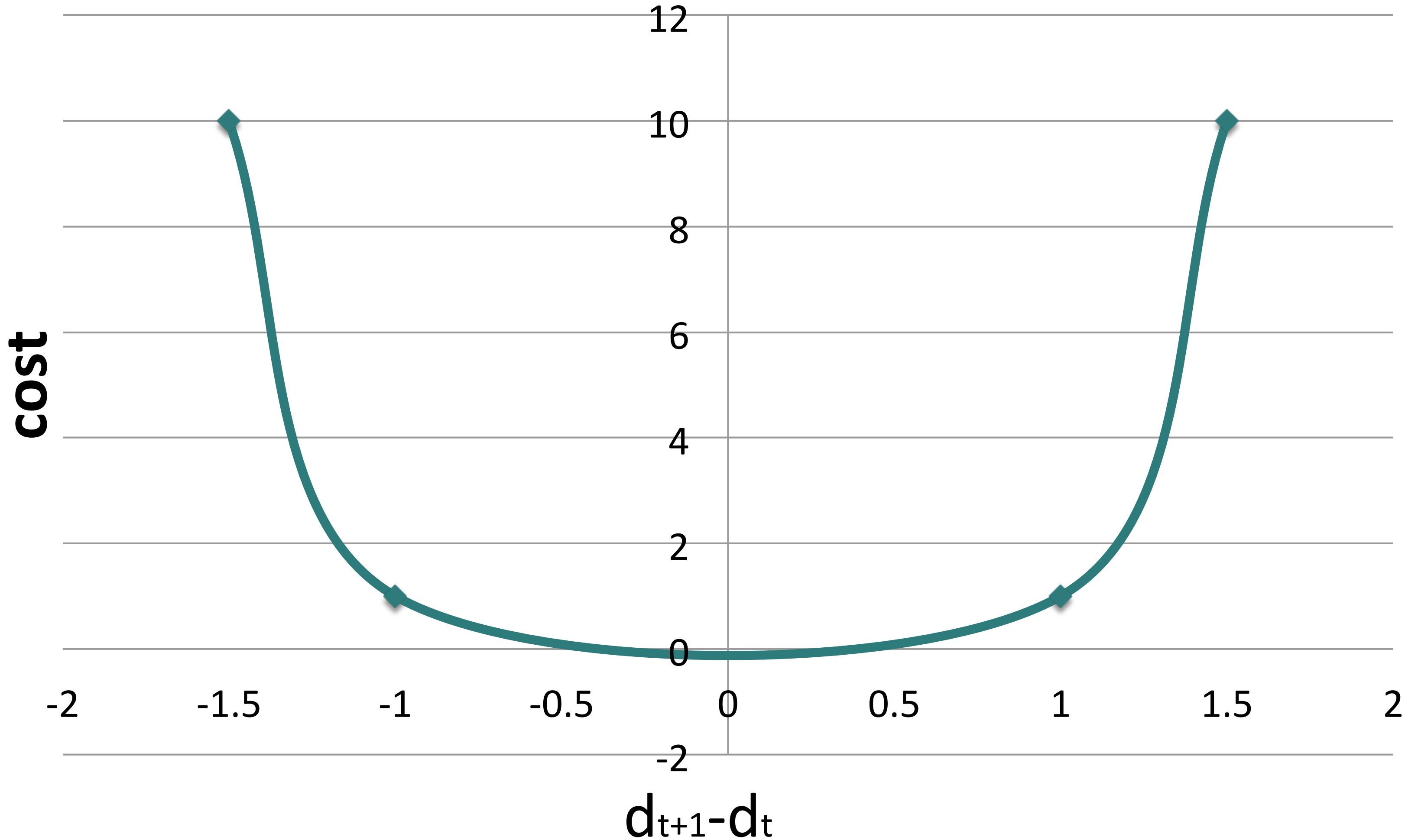
$$F(\theta) = \sum_t \mathbf{p}_t^{\top} log(\mathbf{y}_t) + \|\mathbf{d}_T\| + \sum_{t=0}^{T-1} \eta \xi^{\|\mathbf{d}_{t+1} - \mathbf{d}_t\|}$$

- 1st term : Log-likelihood
- 2nd term: make sure rendering all the information needed
- 3rd term: close only one gate each time step.



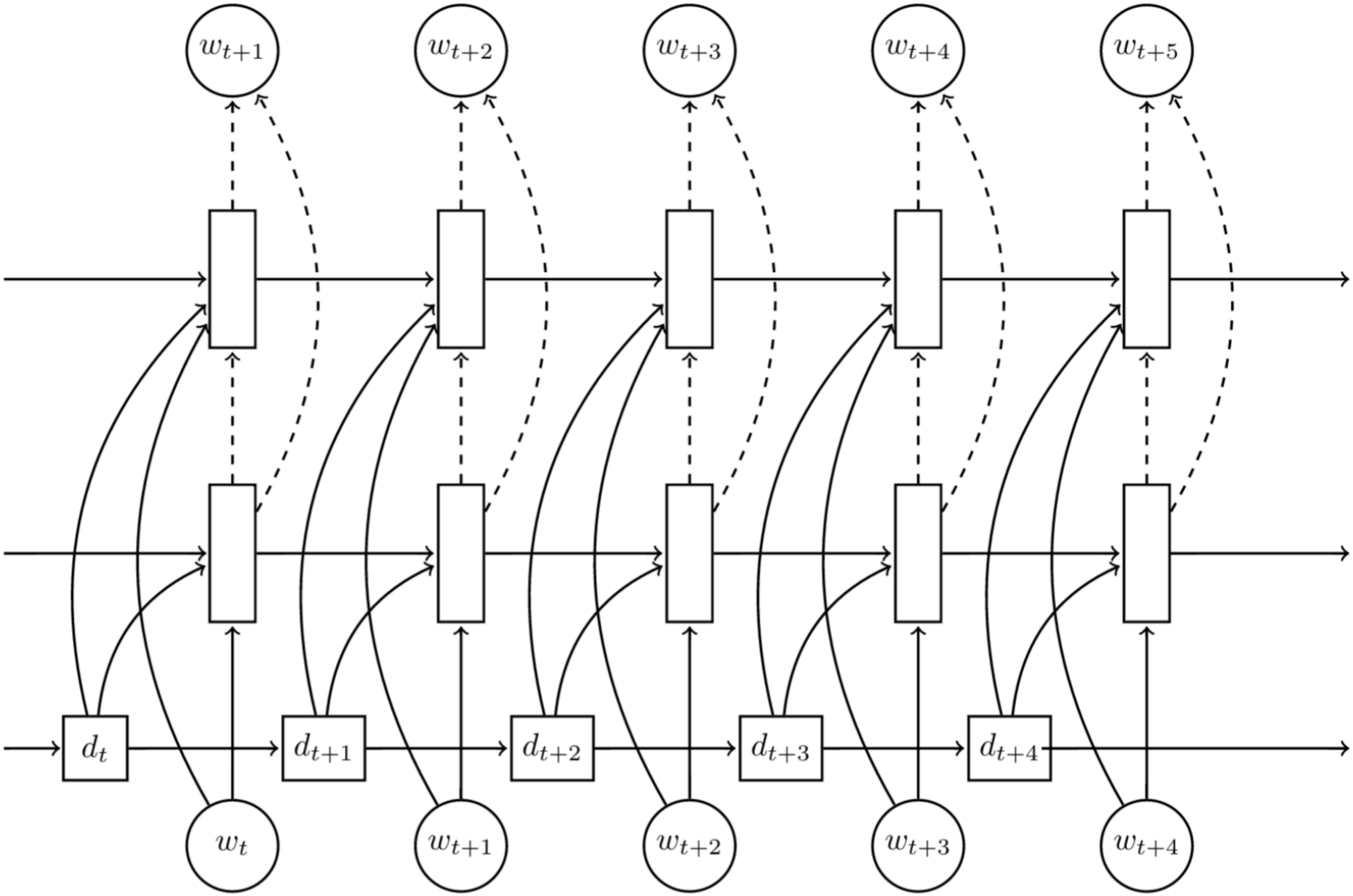
(Hochreiter and Schmidhuber, 1997)

Intuition behind the 3rd term



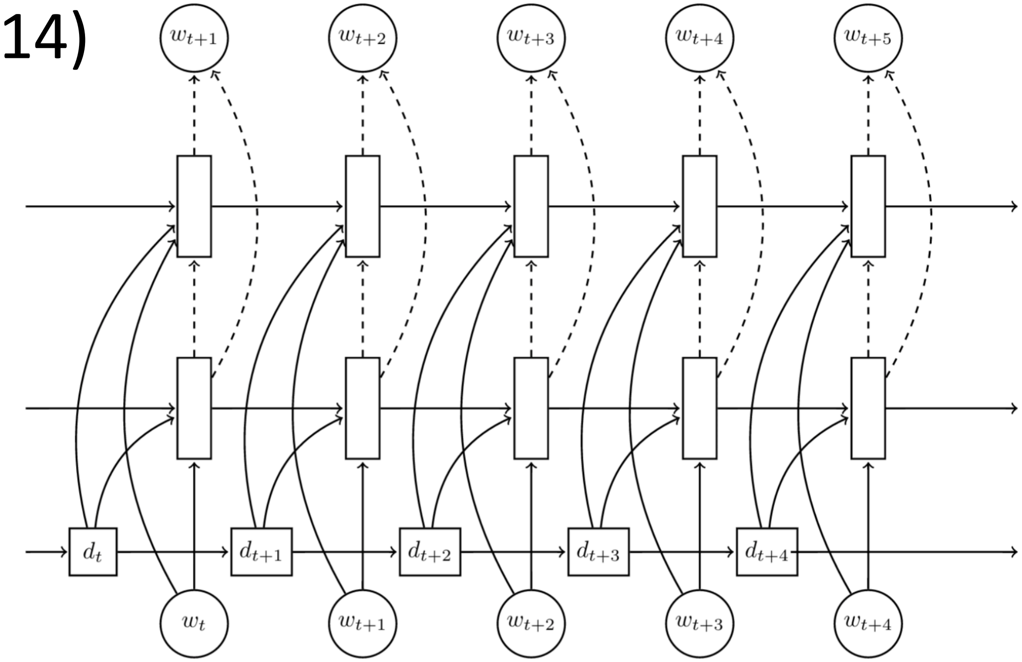
$\eta = 0.01, \xi = 100$

Deep Architecture



Deep Architecture

- Techniques applied
  - Skip connection
    (Graves et al 2013)
  - RNN dropout
    (Srivastava et al 2014)



Outline

- Intro
- Semantically Conditioned LSTM
  - **Experiments**
- Domain adaptation for NLG
- Conclusion

Setup

- Data collection:
  - SFX restaurant/hotel domains

Ontologies

	SF Restaurant	SF Hotel
act type	inform, inform_only, reject, confirm, select, request, reqmore, goodbye	
shared	name, type, *pricerange, price, phone, address, postcode, *area, *near	
specific	*food *goodformeal ***kids-allowed**	***hasinternet** ***acceptscards** ***dogs-allowed**

bold=binary slots, *=slots can take "don't care" value

Setup

- Data collection:
  - SFX restaurant/hotel domains
  - Workers recruited from Amazon MT.
  - Asked to generate system responses given a DA.
  - Result in ~5.1K utterances, 228/164 distinct acts.
- Training: BPTT, L2 reg, SGD w/ early stopping.

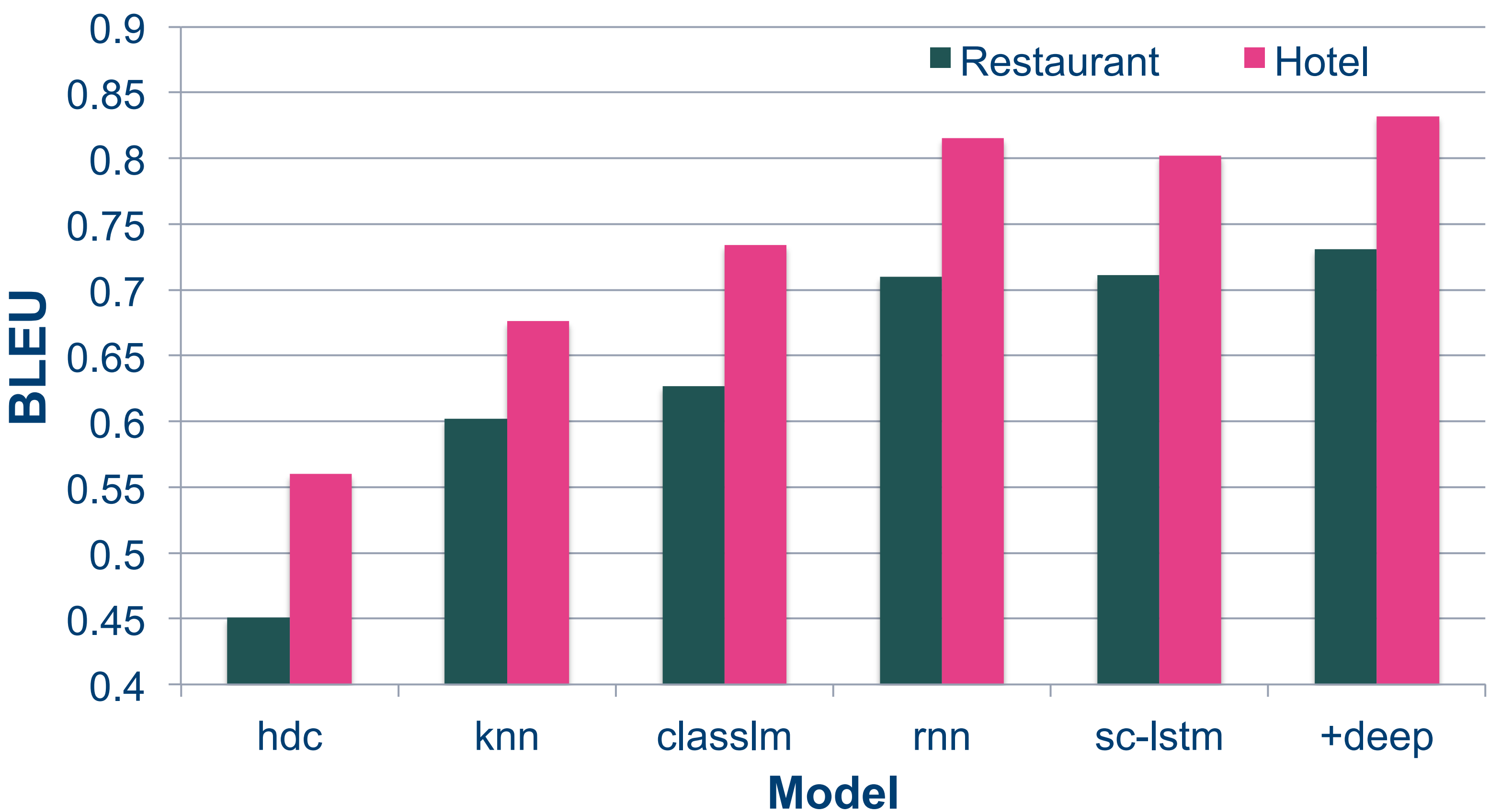
  train/valid/test: 3/1/1, data up-sampling

Available at : https://www.repository.cam.ac.uk/handle/1810/251304

Corpus-based Evaluation

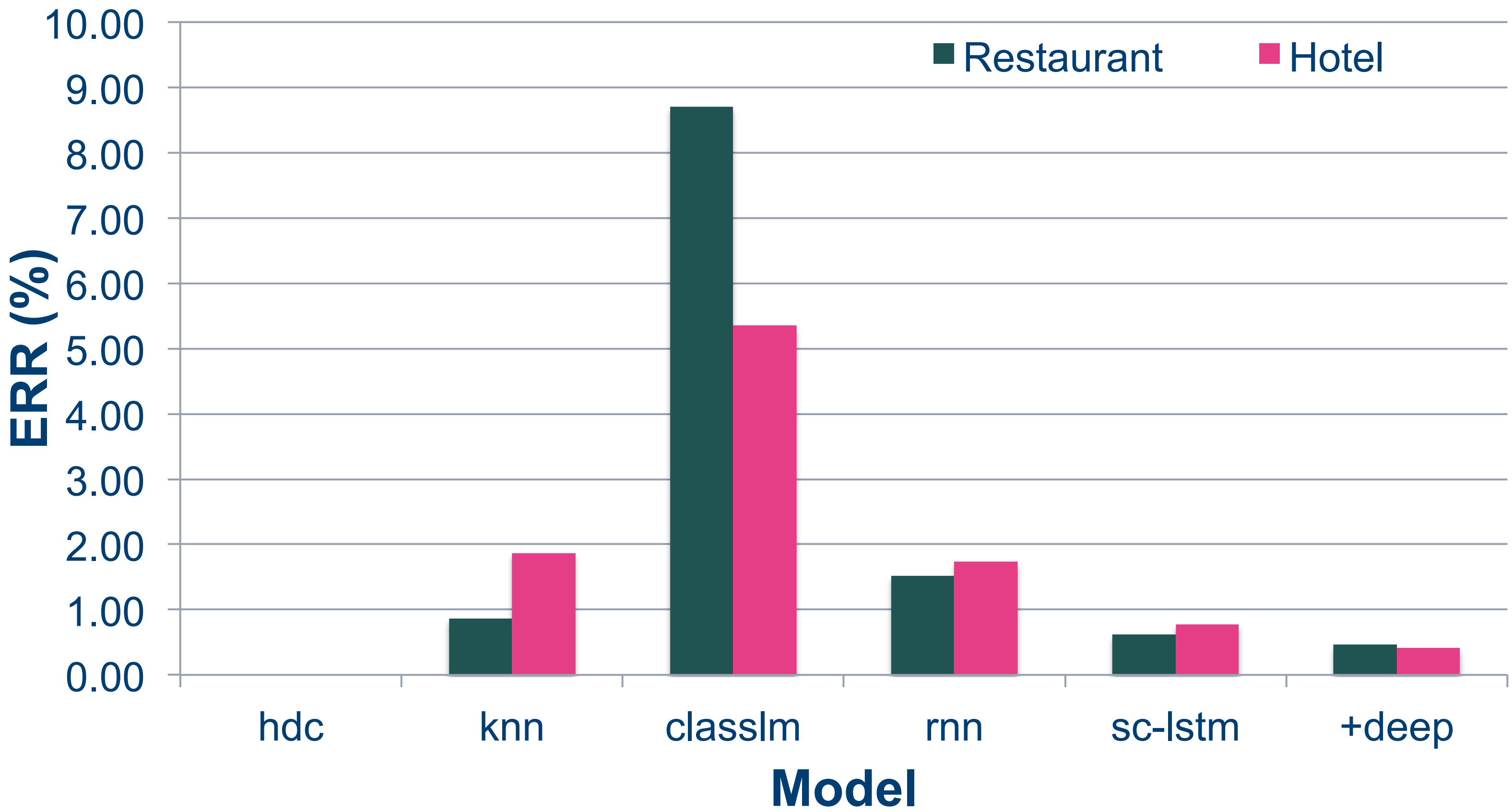
- Test set: ~1K utterances each domain
- Metrics: BLEU-4 (against multiple references), ERR(slot error rates)
- Averaged over 5 random initialised networks.
- Over-gen 20, evaluate on top-5
- Models compared:
  - handcrafted generator (hdc)
  - kNN example-based generator (kNN)
  - class-based LM generator (classlm, O&R 2000)
  - heuristic gated rnn-based generator (rnn, Wen et al 2015)

Corpus-based Evaluation



Selection scheme : 5/20

Corpus-based Evaluation



Selection scheme : 5/20

Generated Examples

inform_only(name='red victorian bed breakfast',accepts_credit_cards='yes',near='haight',has_internet='yes')
red victorian bed breakfast is the only hotel near haight and accepts credit cards and has internet .
red victorian bed breakfast is the only hotel near haight and has internet and accepts credit cards .
red victorian bed breakfast is the only hotel near haight that accept credit cards and offers internet .
the red victorian bed breakfast has internet and near haight , it does accept credit cards .
the red victorian bed breakfast is the only hotel near haight that accepts credit cards , and offers internet .

Generated Examples

inform_no_match(area=tenderloin)

there are no restaurants in the tenderloin area .

there are 0 restaurants in the tenderloin area .

unfortunately there are 0 restaurants in the tenderloin area .

i could not find any restaurants in tenderloin .

there are no restaurants in tenderloin .

Generated Examples

?request(food)
what type of food are you looking for ?
what type of food would you like ?
what kind of food are you looking for ?
what type of food are you interested in ?
what kind of food would you like ?

Human Evaluation

- Setup
  - Judges (~60) recruited from Amazon MT.
  - Asked to evaluate two system responses pairwise.
  - Comparing *classlm*, *rnn*, *sc-lstm*, and *+deep*
- Metrics:
  - Informativeness, Naturalness (rating out of 3)
  - Preference

Human Evaluation

Method	Informativeness	Naturalness
+deep	2.58	**2.51**
sc-lstm	**2.59**	2.50
rnn	2.53	2.42^{*}
classlm	2.46^{**}	2.45

$^{*}p < 0.05$ $^{**}p < 0.005$

Human Evaluation

Pref.%	classlm	rnn	sc-lstm	+deep
classlm	-	46.0	40.9**	37.7**
rnn	54.0	-	43.0	35.7*
sc-lstm	59.1*	57	-	47.6
+deep	62.3**	64.3**	52.4	-

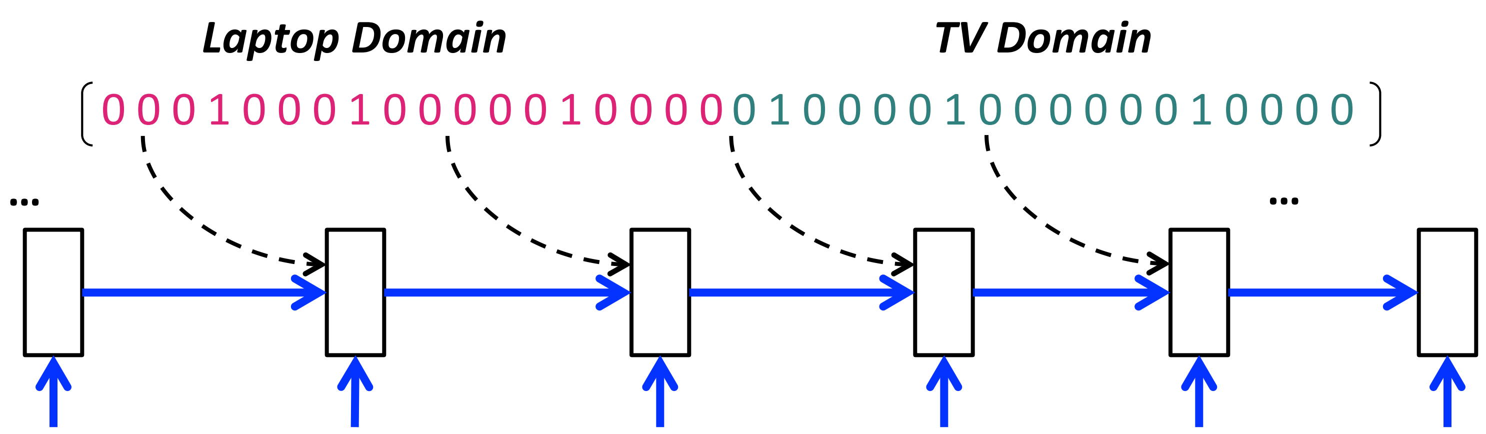
$^{*}p < 0.05$ $^{**}p < 0.005$

Outline

- Intro
- Semantically Conditioned LSTM
- **Domain adaptation for NLG**
  - Data counterfeiting – model initialisation
  - Discriminative training – better fine-tuning
- Conclusion

Domain Adaptation

- Adaptation for NN?
  - Continue to train the model on adaptation dataset
- Parameters are shared on LM part of the network
  - But not for the DA weights
  - New slot-value pairs can only be learned from scratch



Data counterfeiting

- Produce pseudo target domain data by replacing source domain slot-values pairs with target domains slot-value pairs.
- Procedure:

An example realisation in laptop (source) domain:

Zeus 19	is	a	heavy	laptop	with	a	500GB	memory
delexicalisation ⇩								
<NAME-value>	is	a	<WEIGHT-value>	<TYPE-value>	with	a	<MEMEORY-value>	<MEMORY-slot>
counterfeiting ⇩								
<NAME-value>	is	a	<FAMILY-value>	<TYPE-value>	with	a	<SCREEN-value>	<SCREEN-slot>

A possible realisation in TV (target) domain:

Apollo 73	is	a	U76	television	with	a	29-inch	screen

Data counterfeiting

- Choice of target domain slots?
  - The realisation should be similar to the source one.
  - Simple case: based on their functional class.
    - Informable, requestable, and binary slots.
  - Example:

	Laptop	**Television**
Informable	family, price_range, battery_rating,…	family, price_range, screen_size_range,…
Requestable	price, memory,…	price, resolution,…
Binary	is_for_business	has_usb_port

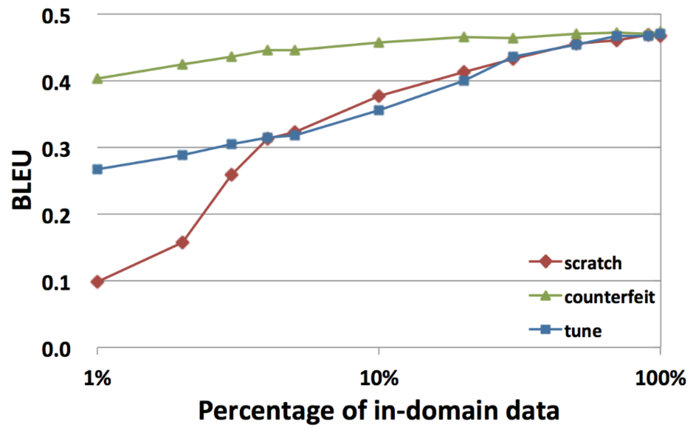
Laptop/TV dataset

- A more difficult dataset than restaurant/hotel
  - Permutate all possible DAs, ~13K/7K
  - Only 1 example utterance for each DA

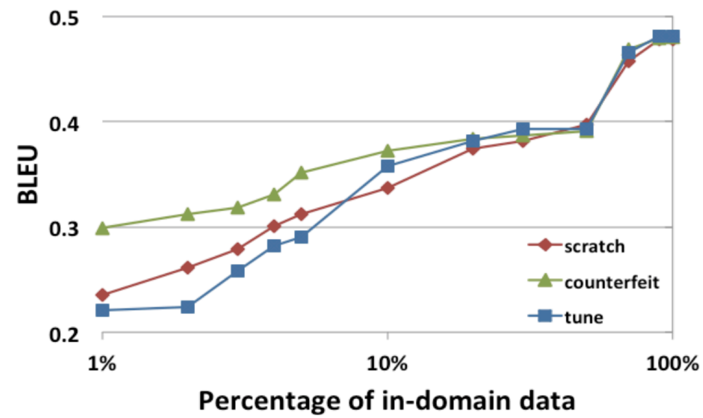
	Laptop	Television
informable slots	family, *pricerange, batteryrating, driverange, weightrange, **isforbusinesscomputing**	family, *pricerange, screensizerange, ecorating, hdmiport, **hasusbport**
requestable slots	*name, *type, *price, warranty, battery, design, dimension, utility, weight, platform, memory, drive, processor	*name, *type, *price, resolution, powerconsumption, accessories, color, screensize, audio
act type	*inform, *inform_only_match, *inform_on_match, inform_all, *inform_count, inform_no_info, *recommend, compare, *select, suggest, *confirm, *request, *request_more, *goodbye	

bold=binary slots, *=overlap with SF Restaurant and Hotel domains, all *informable slots* can take ”dontcare” value

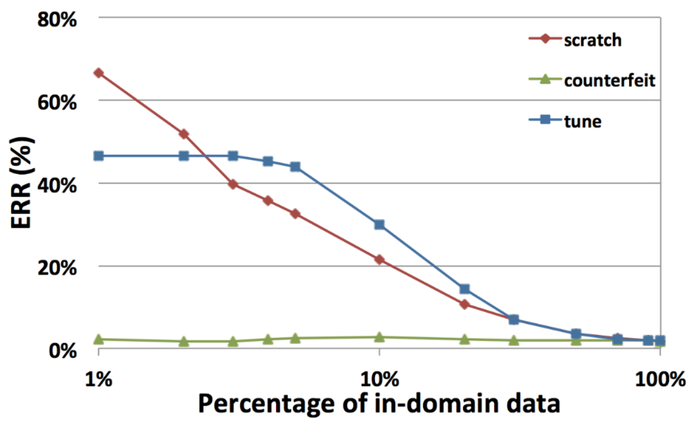
Data counterfeiting - Results



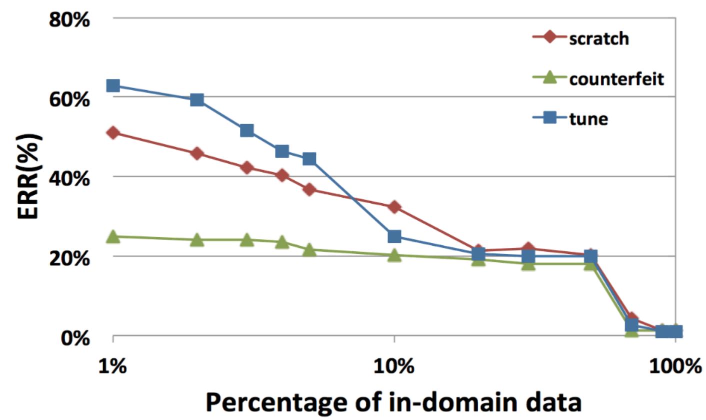
(a) BLEU score curve



(a) BLEU score curve



(b) Slot error rate curve



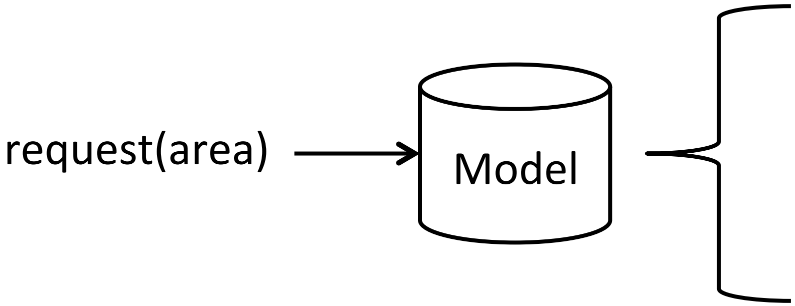
(b) Slot error rate curve

Laptop 2 TV

Restaurant+Hotel 2 Laptop+TV

Discriminative Training

- Explore model capacity and correct it.



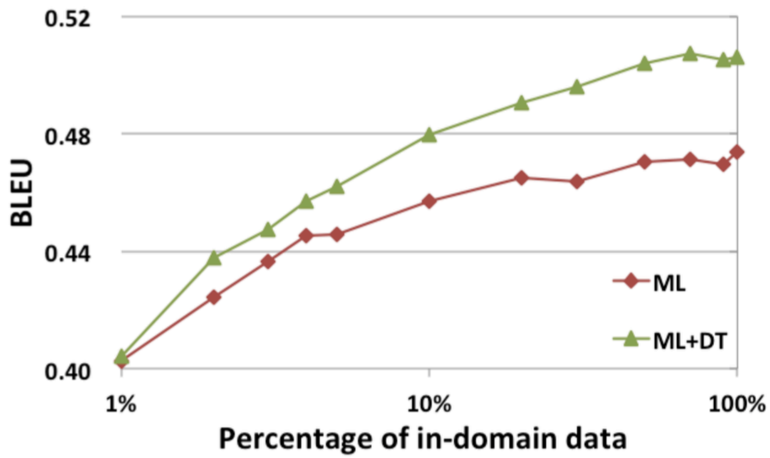
candidates	*score*
What area do you want?	0.9
What food type do you want?	0.2
Do you want north area?	0.1
Do you have any area in mind?	0.8
What part of town do you want?	1.0

- DT cost function:

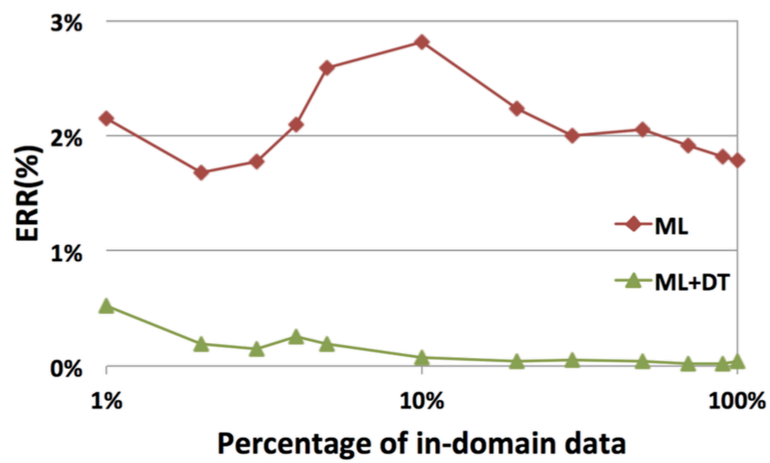
$$F(\theta) = -\mathbb{E}[L(\theta)]$$
$$= -\sum_{\Omega \in Gen(d_i)} p_\theta(\Omega|d_i) L(\Omega, \Omega_i)$$

Ω: candidate sentence
Ωi: reference sentence
di: dialogue act
L(.): scoring function

Discriminative Training - Results



(a) Effect of DT on BLEU



(b) Effect of DT on slot error rate

Human Evaluation

Method	TV to Laptop Info.	TV to Laptop Nat.	laptop to TV Info.	laptop to TV Nat.
scrALL	2.64	2.37	2.54	2.36
DT-10%	**2.52****	**2.25****	**2.51**	2.19**
ML-10%	2.51**	2.22**	2.45**	**2.22****
scr-10%	2.24**	2.03**	2.00**	1.92**

* $p < 0.05$, ** $p < 0.005$

- scrALL : train from scratch with 100% ID data.
- scr-10% : train from scratch with 10% ID data.
- ML-10% : data counterfeiting + ML training on 10% ID data.
- DT-10% : data counterfeiting + DT training on 10% ID data.

Outline

- Intro
- Semantically Conditioned LSTM
- Domain adaptation for NLG
- **Conclusion**

Conclusion

- NLG can be learned N2N from data.
  - Learn LM & slot gating control signal jointly
  - Corpus-based/Human evaluation.
  - More colloquial, more scalable.

- Domain Extension
  - Data counterfeiting facilitates domain adaptation.
  - Discriminative training can further improve.

Papers

- Tsung-Hsien Wen, Milica Gasic , Dongho Kim, Nikola Mrksic, Pei-Hao Su, David Vandyke, and Steve Young. Stochastic language generation in dialogue using recurrent neural networks with convolutional sentence reranking. In *Proceedings of SIGdial 2015*.
- Tsung-Hsien Wen, Milica Gasic , Nikola Mrksic, Pei-Hao Su, David Vandyke, and Steve Young. Semantically Conditioned LSTM-based Natural Language Generation for Spoken Dialogue Systems. In *Proceedings of EMNLP 2015*.
- Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Lina M.R. Barahona, Pei-Hao Su, David Vandyke, and Steve Young. Muti-domain Neural Language Generation for Spoken Dialogue Systems. Submitting to NAACL 2016.

Selected References

- Amanda Stent, Matthew Marge, and Mohit Singhai. 2005. Evaluating evaluation methods for generation in the presence of variation. In Proceedings of CICLing 2005.
- Alice H. Oh and Alexander I. Rudnicky. 2000. Stochastic language generation for spoken dialogue systems. In Proceedings of the 2000 ANLP/NAACL Workshop on Conversational Systems.
- Tomas Mikolov, Martin Karafit, Lukas Burget, Jan Cernocky, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. *In Proceedings on InterSpeech*.
- Sepp Hochreiter and Jurgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*.

UNIVERSITY OF CAMBRIDGE

Thank you! Questions?

Tsung-Hsien Wen is supported by a studentship funded by Toshiba Research Europe Ltd, Cambridge Research Laboratory