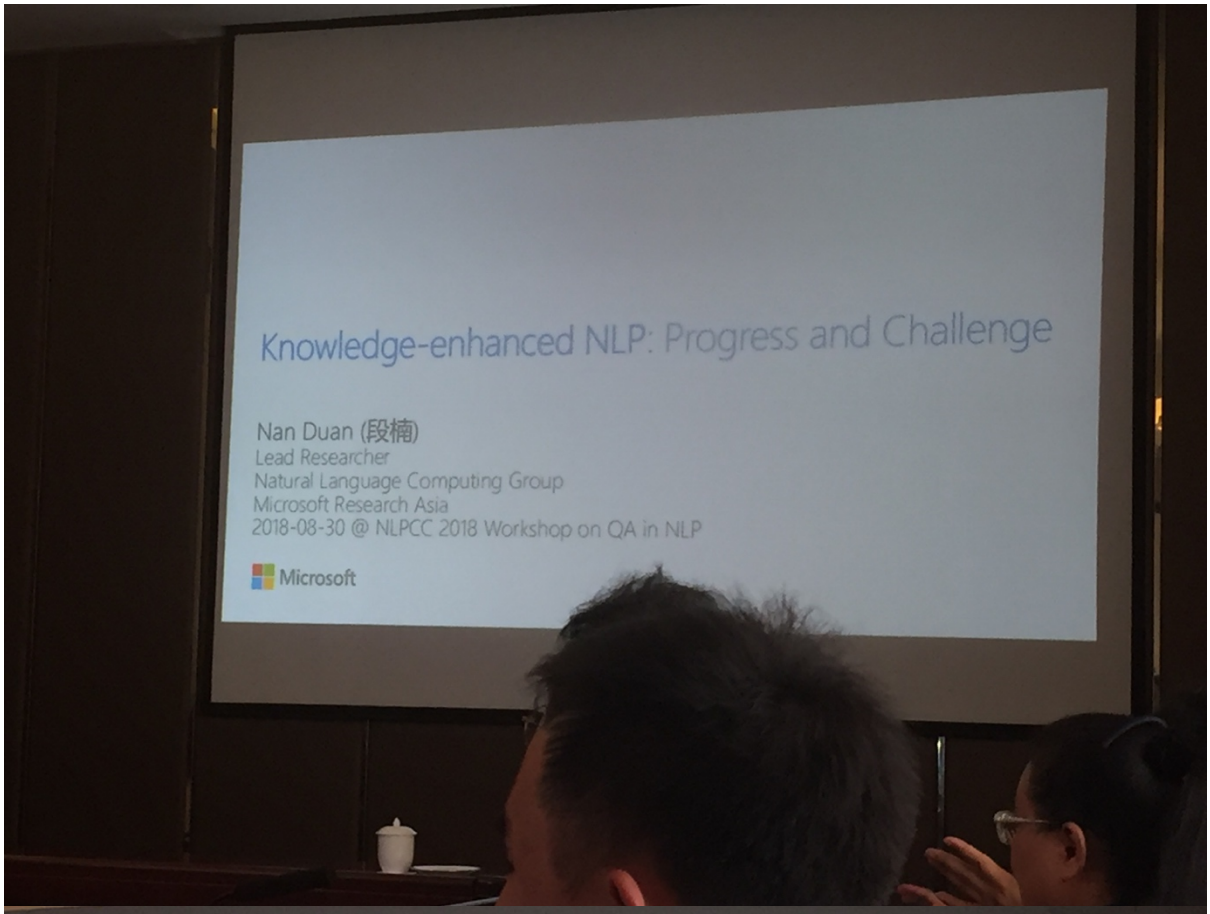
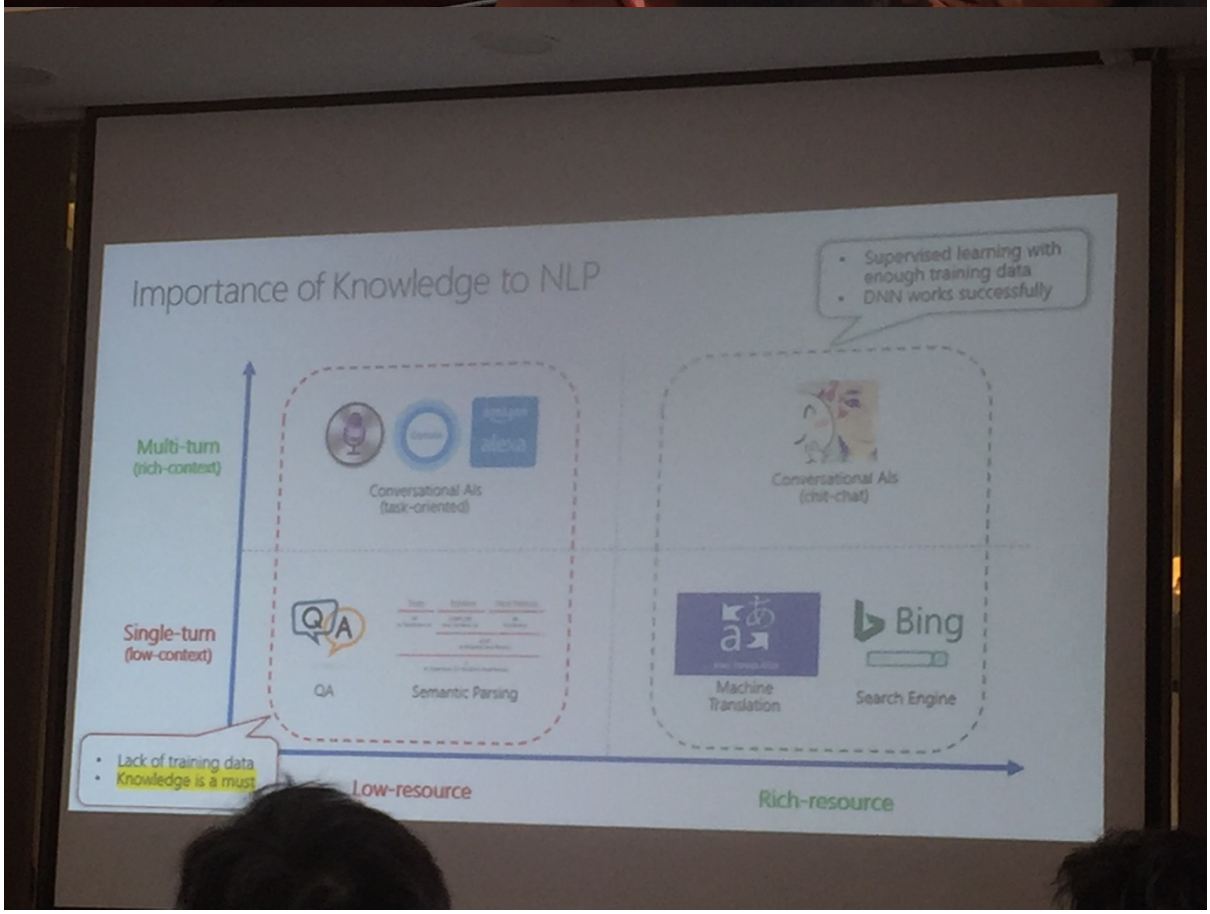
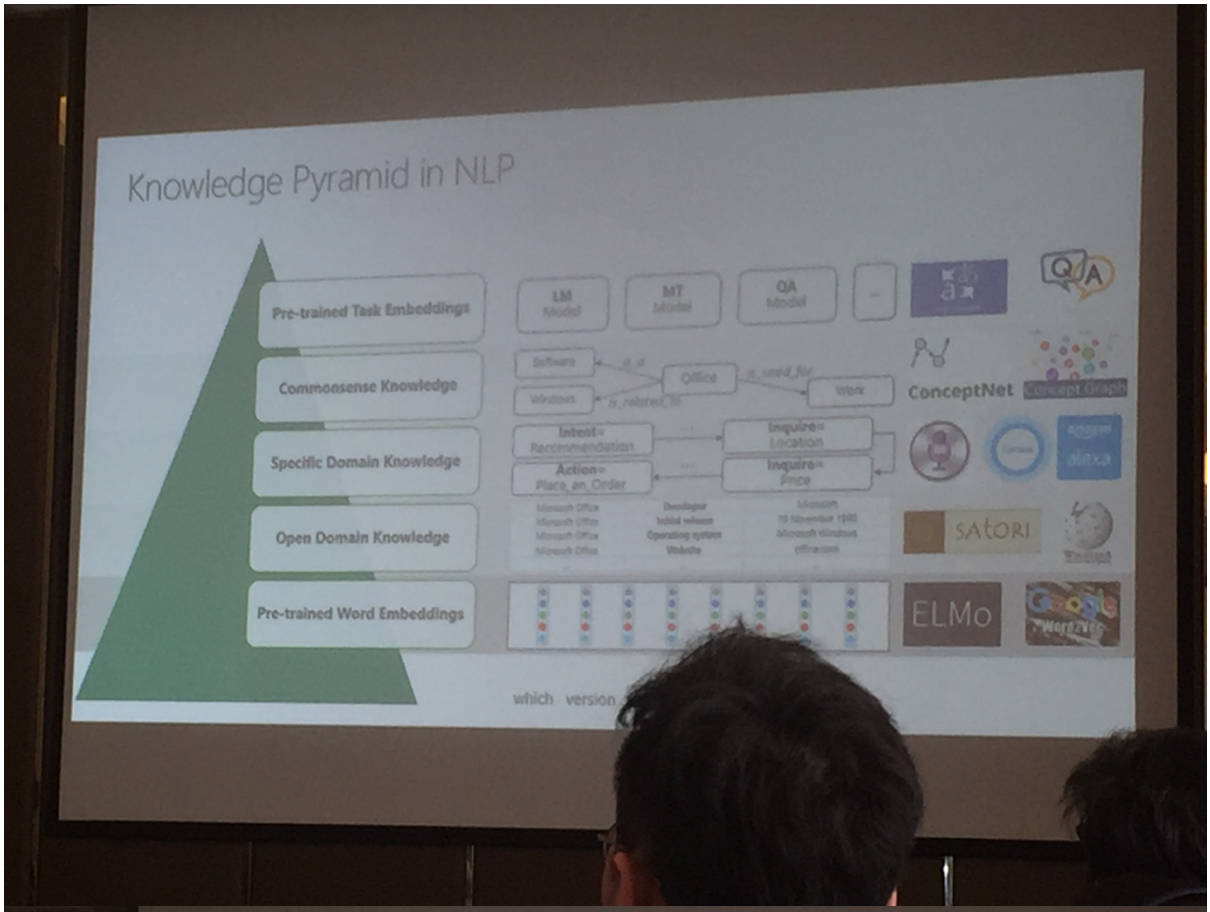
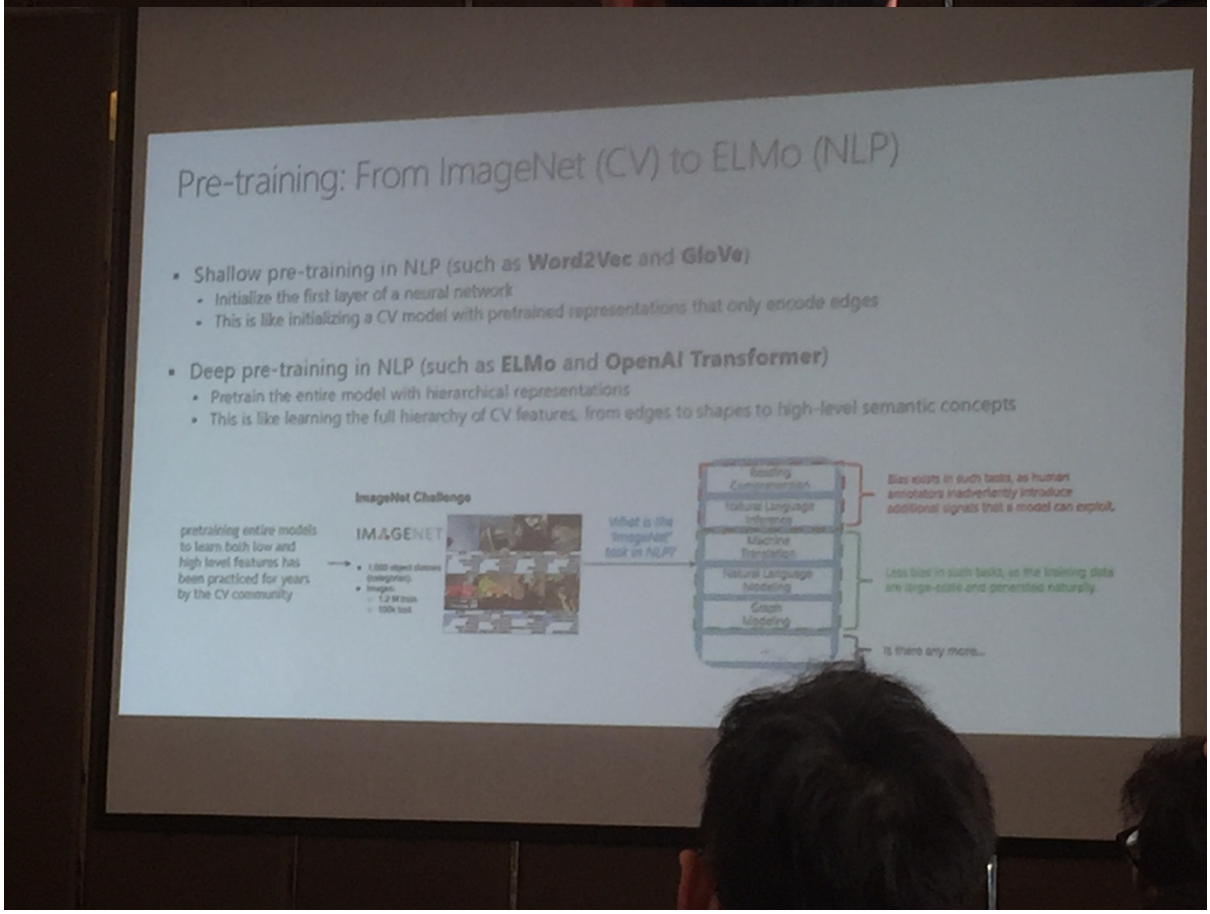
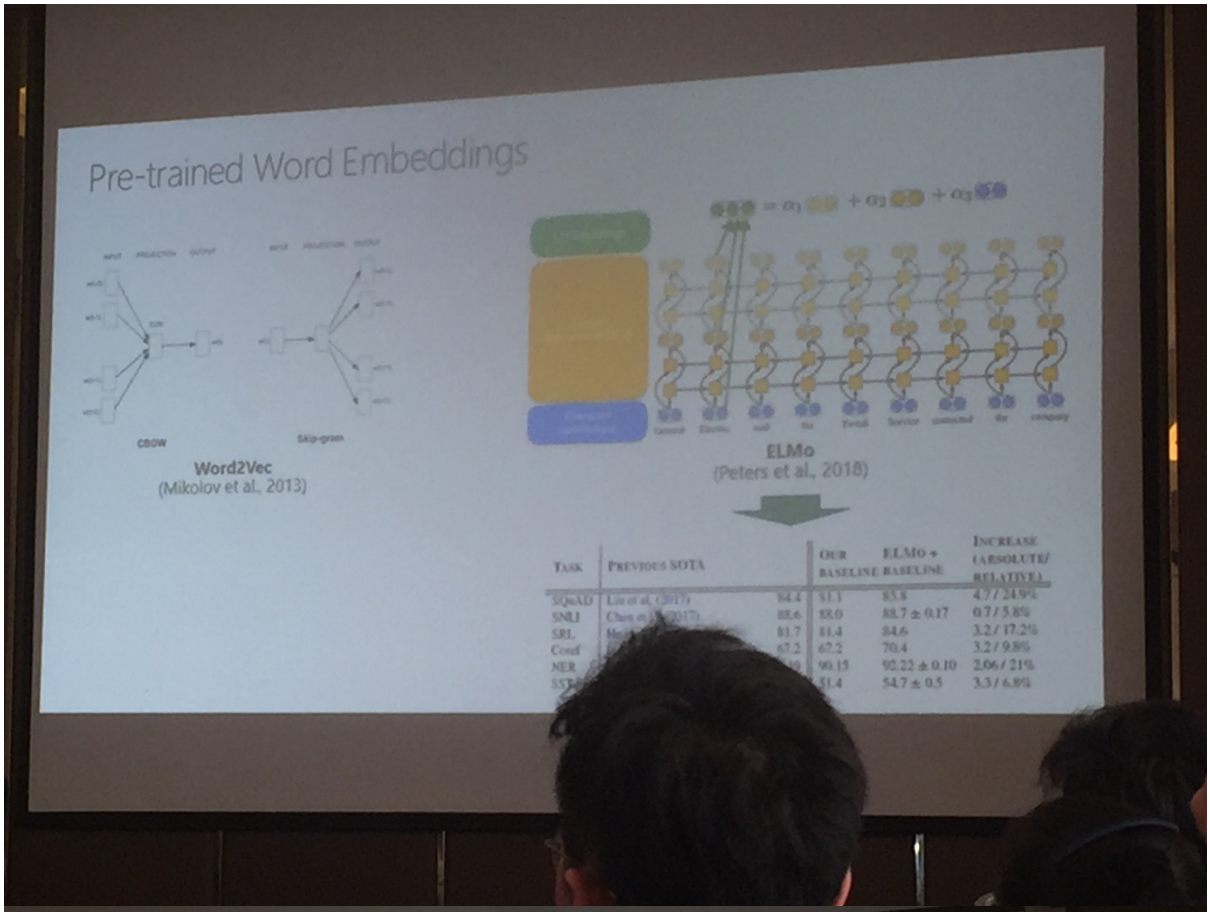
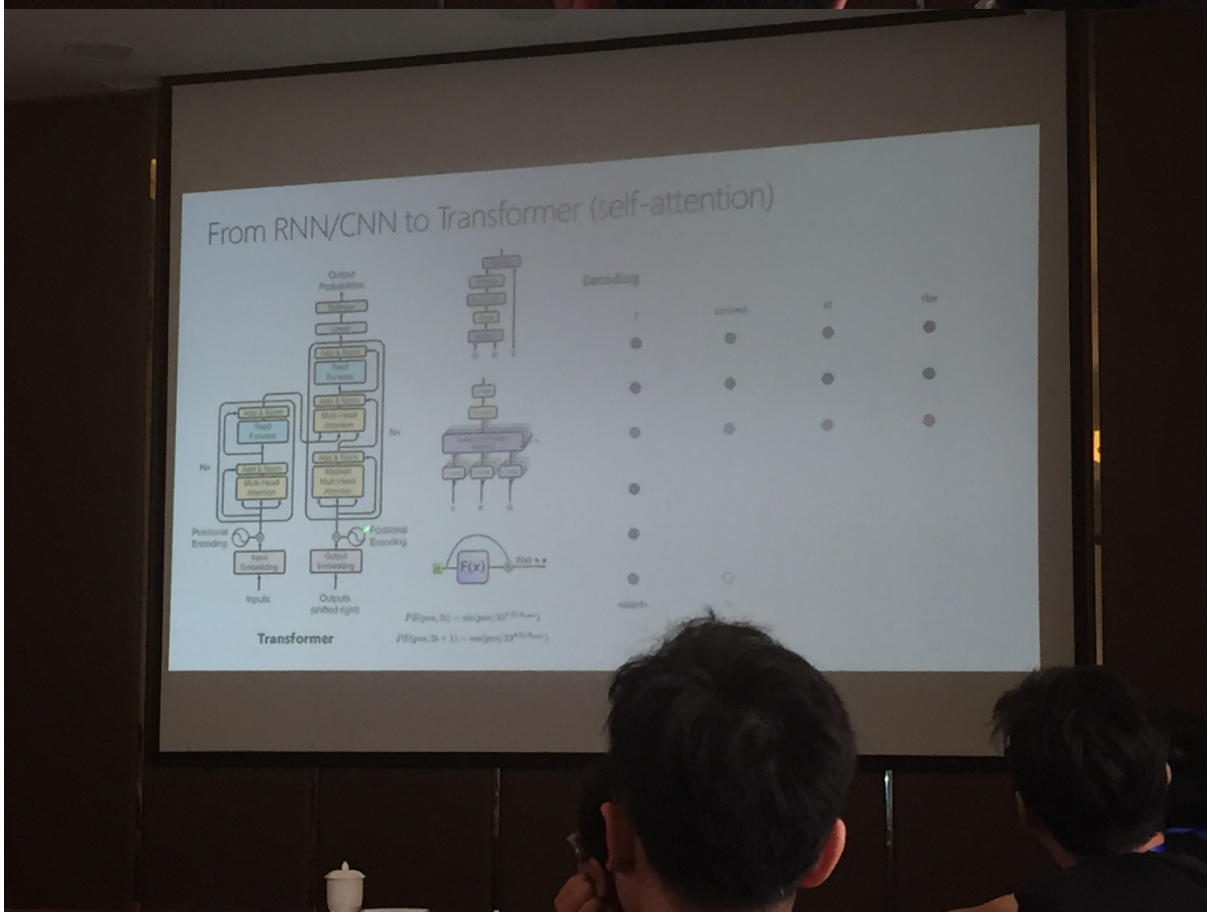

Knowledge-enhanced NLP: Progress and Challenge
Nan Duan (段楠)
Lead Researcher
Natural Language Computing Group
Microsoft Research Asia
2018-08-30 @ NLPCC 2018 Workshop on QA in NLP
Microsoft


Importance of Knowledge to NLP
Supervised learning with enough training data
DNN works successfully
Multi-turn (rich-context)
Single-turn (low-context)
Conversational AIs (task-oriented)
Conversational AIs (chit-chat)
QA
Semantic Parsing
Machine Translation
Bing
Search Engine
Lack of training data
Knowledge is a must
Low-resource
Rich-resource

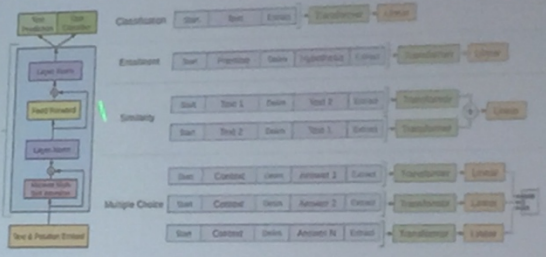

Knowledge Pyramid in NLP
Pre-trained Task Embeddings
Commonsense Knowledge
Specific Domain Knowledge
Open Domain Knowledge
Pre-trained Word Embeddings
LM Model
MT Model
QA Model
ConceptNet
Concept Graph
alexa
SATORI
ELMo
which version


Pre-training: From ImageNet (CV) to ELMo (NLP)
Shallow pre-training in NLP (such as Word2Vec and GloVe)
Initialize the first layer of a neural network
This is like initializing a CV model with pretrained representations that only encode edges
Deep pre-training in NLP (such as ELMo and OpenAI Transformer)
Pretrain the entire model with hierarchical representations
This is like learning the full hierarchy of CV features, from edges to shapes to high-level semantic concepts
pretraining entire models to learn both low and high level features has been practiced for years by the CV community
ImageNet Challenge
IMAGENET
What is the "ImageNet" task in NLP?
Reading Comprehension
Natural Language Inference
Machine Translation
Natural Language Modeling
Graph Modeling
Bias exists in such tasks, as human annotators inadvertently introduce additional signals that a model can exploit.
Less bias in such tasks, as the training data are large-scale and generated naturally.
Is there any more...


Pre-trained Word Embeddings
CBOW
Skip-gram
Word2Vec
(Mikolov et al., 2013)
ELMo
(Peters et al., 2018)
TASK
PREVIOUS SOTA
OUR BASELINE
ELMo + BASELINE
INCREASE (ABSOLUTE/ RELATIVE)
SQuAD
SNLI
SRL
Coref
NER
SST-5
85.8
88.7 ± 0.17
84.6
70.4
92.22 ± 0.10
54.7 ± 0.5
4.7 / 24.9%
0.7 / 5.8%
3.2 / 17.2%
3.2 / 9.8%
2.06 / 21%
3.3 / 6.8%


From RNN/CNN to Transformer (self-attention)
Transformer
Decoding

Unsupervised Pre-training + Supervised Fine-tuning

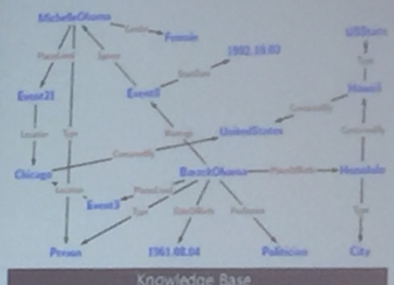


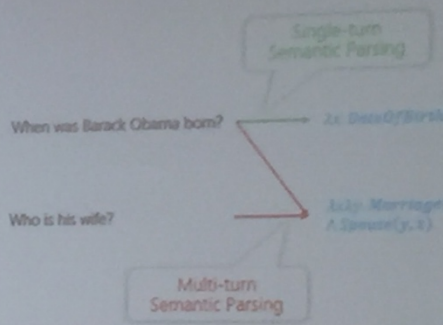
Method	Story Cloze	RACE-m	RACE-h	RACE
val-LS-skip [55]	76.5	-	-	-
Hidden Coherence Model [7]	77.6	-	-	-
Dynamic Fusion Net [67] (9x)	-	55.6	49.4	51.2
BiAttention MRU [59] (9x)	-	60.2	50.3	53.3
Finetuned Transformer LM (ours)	**86.5**	**62.9**	**57.4**	**59.0**

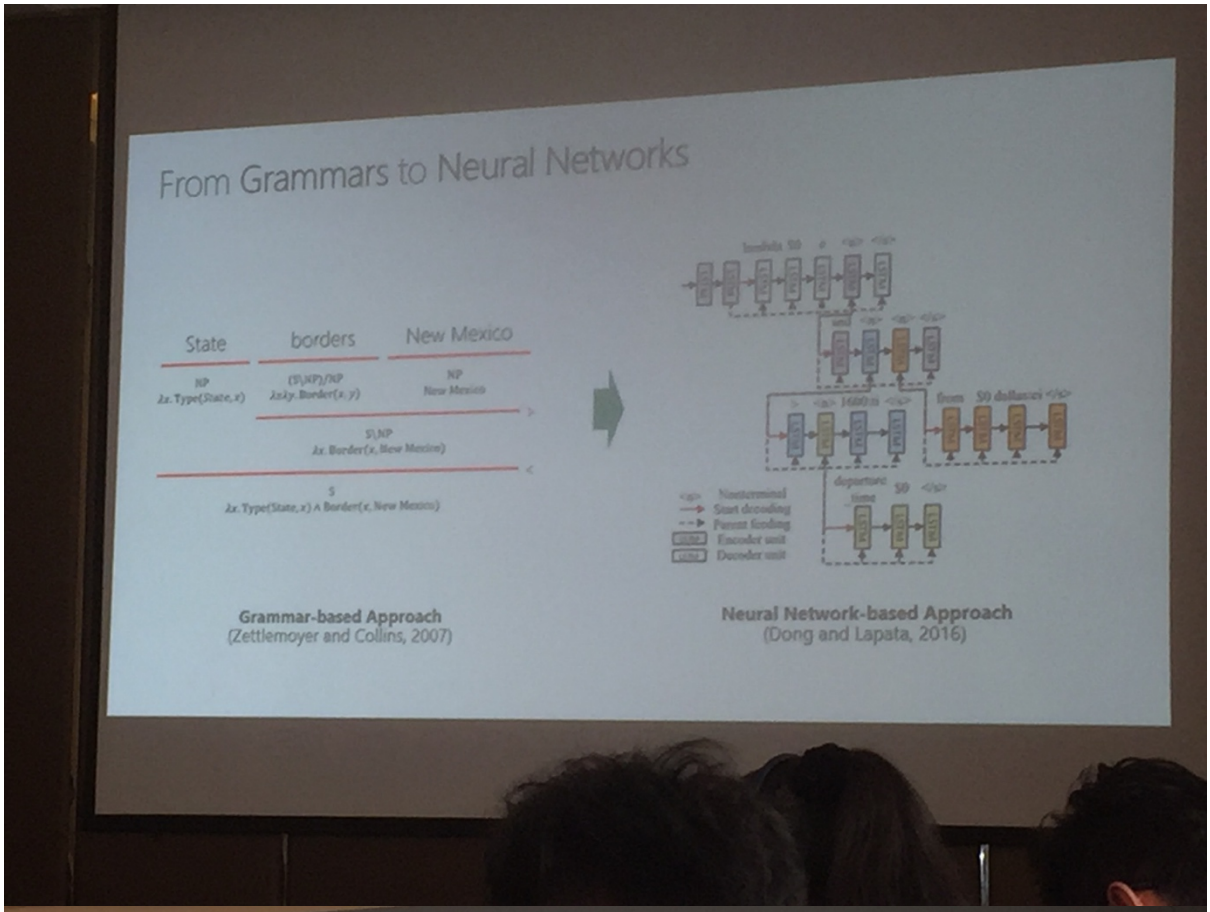
Method	Classification		Semantic Similarity			GLUE
	CoLA (mc)	SST2 (acc)	MRPC (F1)	STSB (pc)	QQP (F1)	
Sparse byte mLSTM [16]	-	**93.2**	-	-	-	-
TF-KLD [23]	-	-	**86.0**	-	-	-
ECNU (mixed ensemble) [60]	-	-	-	**81.0**	-	-
Single-task BiLSTM + ELMo + Attn [64]	35.0	90.2	80.2	55.5	66.1	64.8
Multi-task BiLSTM + ELMo + Attn [64]	18.9	91.6	83.5	72.8	63.3	68.9
Finetuned Transformer LM (ours)	**45.4**	91.3	82.3	**82.0**	**70.3**	**72.8**

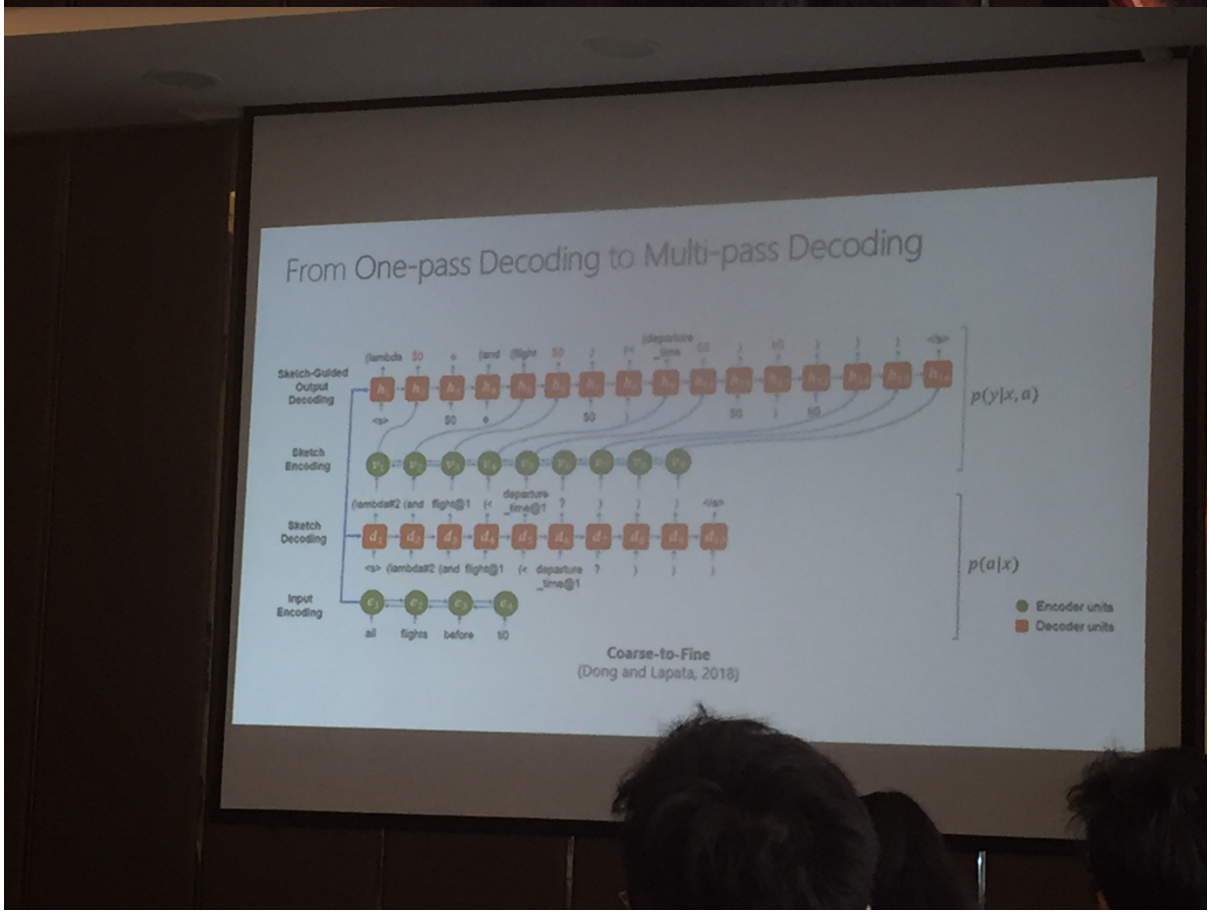
Take Semantic Parsing as an Example

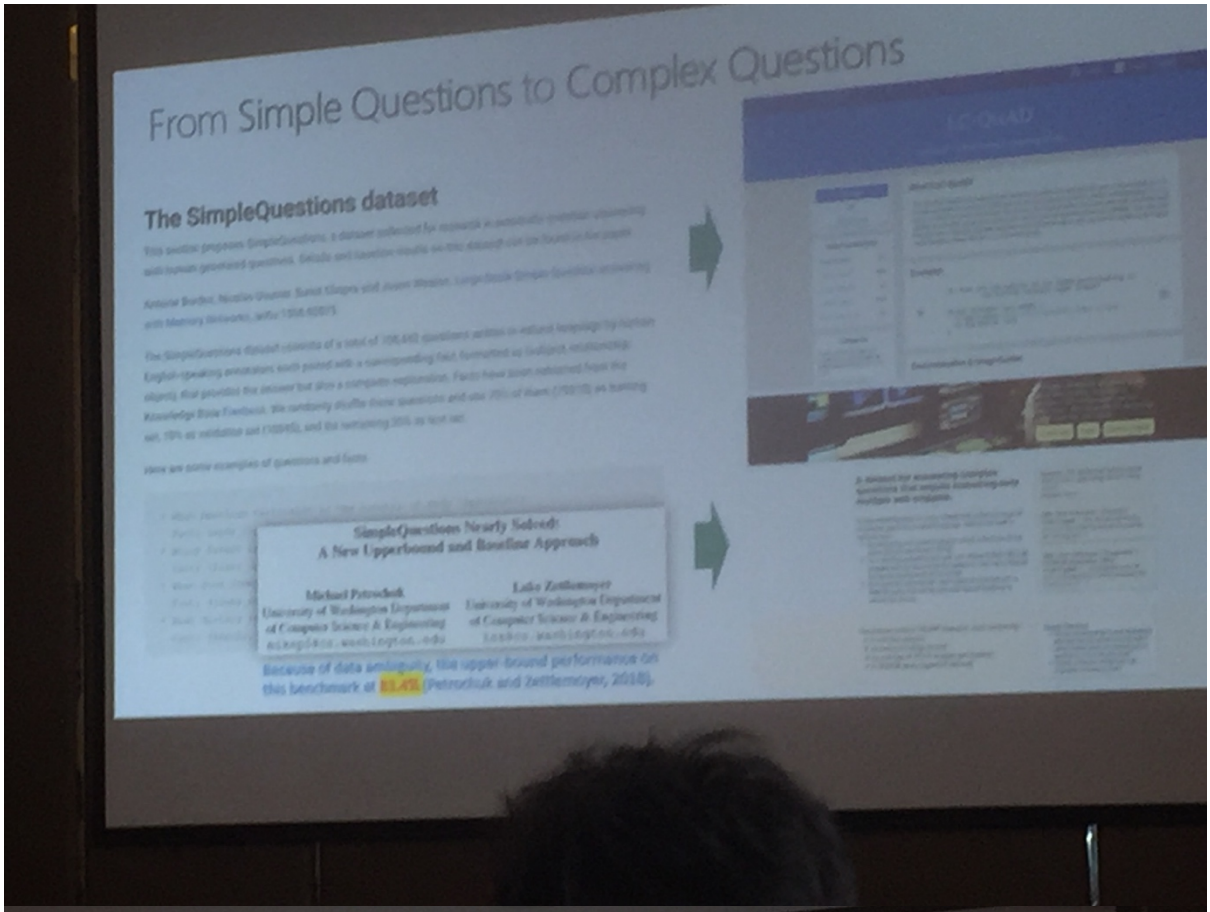
- Convert NL utterances into machine executable LFs based on knowledge base

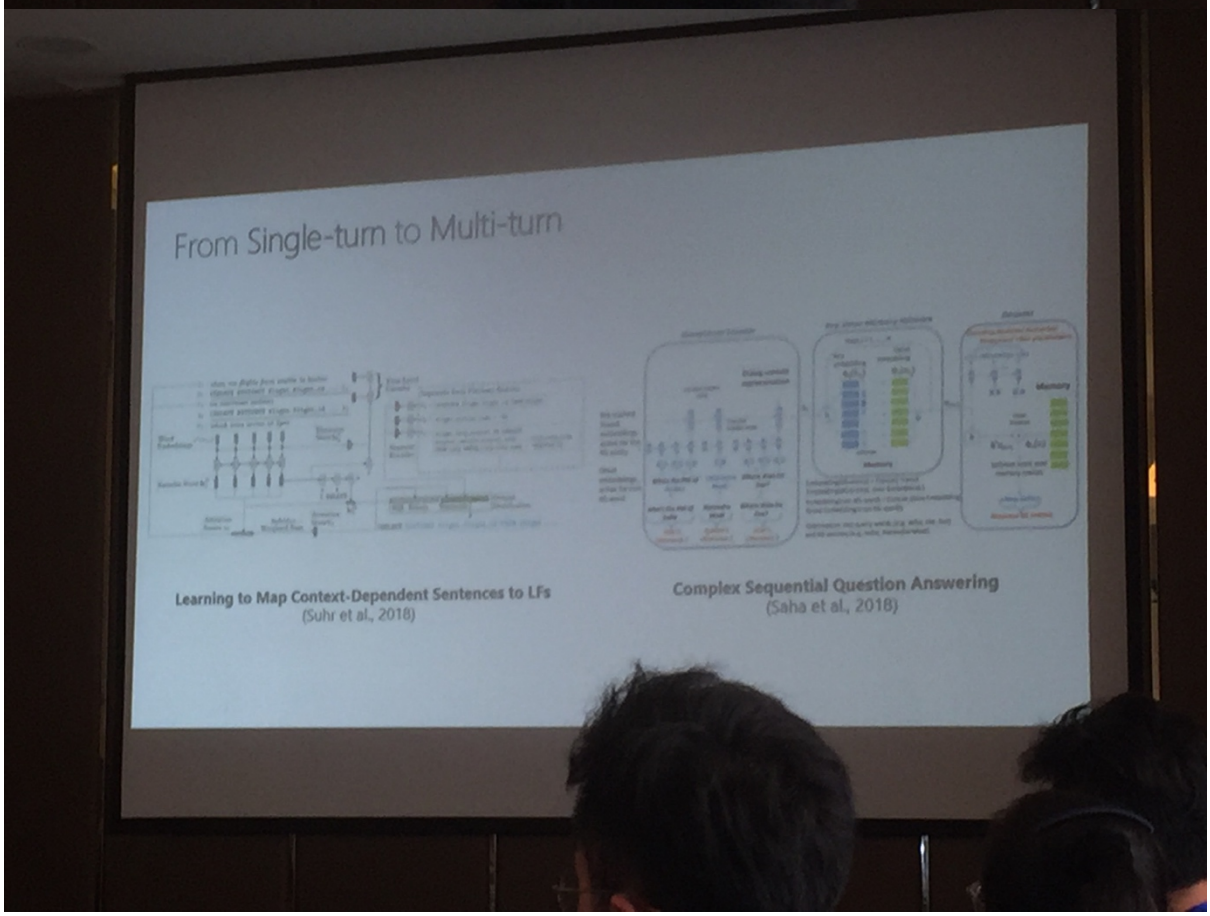




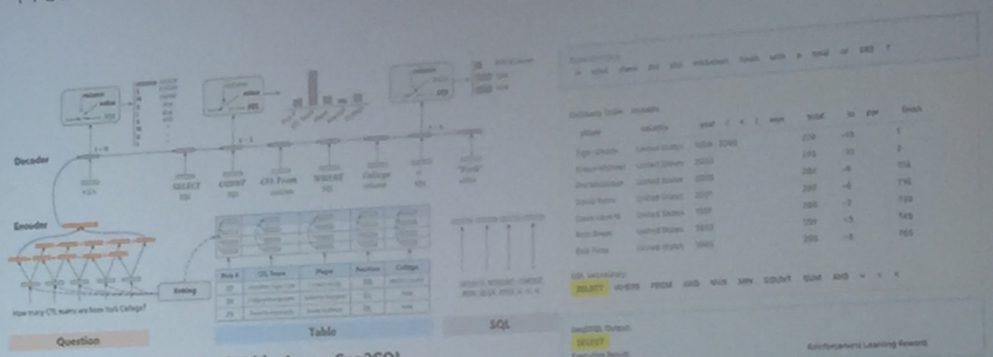

From Grammars to Neural Networks
State
borders
New Mexico
NP
λx. Type(State, x)
(S\NP)/NP
λxλy. Border(x, y)
NP
New Mexico
S\NP
λx. Border(x, New Mexico)
S
λx. Type(State, x) ∧ Border(x, New Mexico)
Grammar-based Approach
(Zettlemoyer and Collins, 2007)
Neural Network-based Approach
(Dong and Lapata, 2016)


From One-pass Decoding to Multi-pass Decoding
Sketch-Guided Output Decoding
Sketch Encoding
Sketch Decoding
Input Encoding
all
flights
before
$0
p(y|x, a)
p(a|x)
Encoder units
Decoder units
Coarse-to-Fine
(Dong and Lapata, 2018)


From Simple Questions to Complex Questions
The SimpleQuestions dataset
SimpleQuestions Nearly Solved:
A New Upperbound and Baseline Approach
Michael Petrochuk
Luke Zettlemoyer
Because of data ambiguity, the upper-bound performance on this benchmark is 83.4% (Petrochuk and Zettlemoyer, 2018).

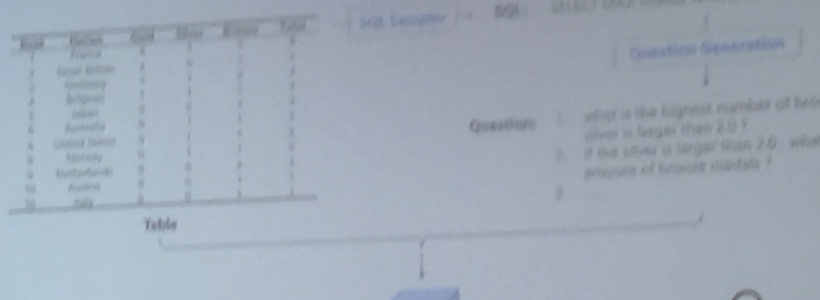

From Single-turn to Multi-turn
Learning to Map Context-Dependent Sentences to LFs
(Suhr et al., 2018)
Complex Sequential Question Answering
(Saha et al., 2018)

From Knowledge Graph to Web Tables



Syntax- and Table-Aware Seq2SQL
(Sun et al., 2018)

From Supervised Learning to Semi-supervised Learning

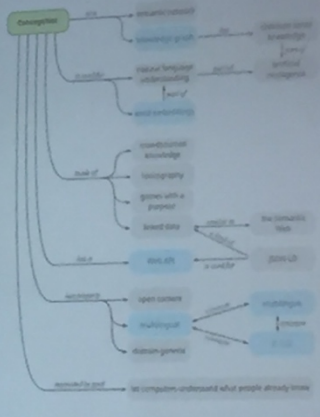


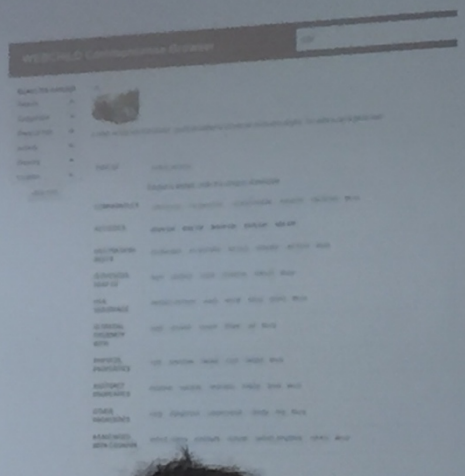
QG Improves Seq2SQL
(Guo et al., 2018)

Conclusion (2)

- Training data is critical to natural language understanding
- Previous work focus on single-turn, single-relation and open domain tasks
- Current work moves to multi-turn, complex-relation and specific tasks
- Take-aways
  - Single-turn questions are almost solved, if we have enough data
  - Multi-turn & complex questions are still challenging

Commonsense Knowledge





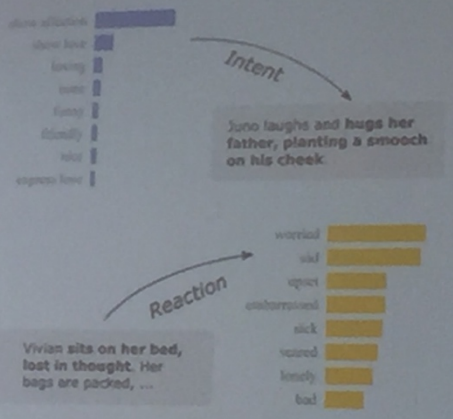
ConceptNet

Intent and Reaction Prediction with Commonsense Knowledge

- Use commonsense knowledge learnt from **supervised data**

PersonX cooks thanksgiving dinner	X's intent	to impress their family
	X's reaction	tired, a sense of belonging
	Y's reaction	impressed
PersonX drags PersonX's feet	X's intent	to avoid doing things
	X's reaction	lazy, bored
	Y's reaction	frustrated, impatient
PersonX reads PersonY's diary	X's intent	to be nosey, know secrets
	X's reaction	guilty, curious
	Y's reaction	angry, violated, betrayed

Event2Mind
(Rashkin et al., 2018)

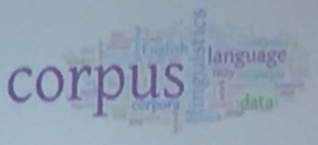


QA and Reasoning with Commonsense Knowledge

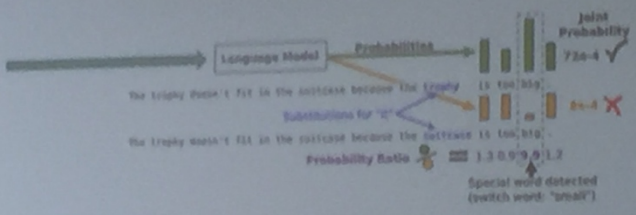
*The service was poor, **but** the food was ...*

- Use commonsense knowledge learnt from **unsupervised data**

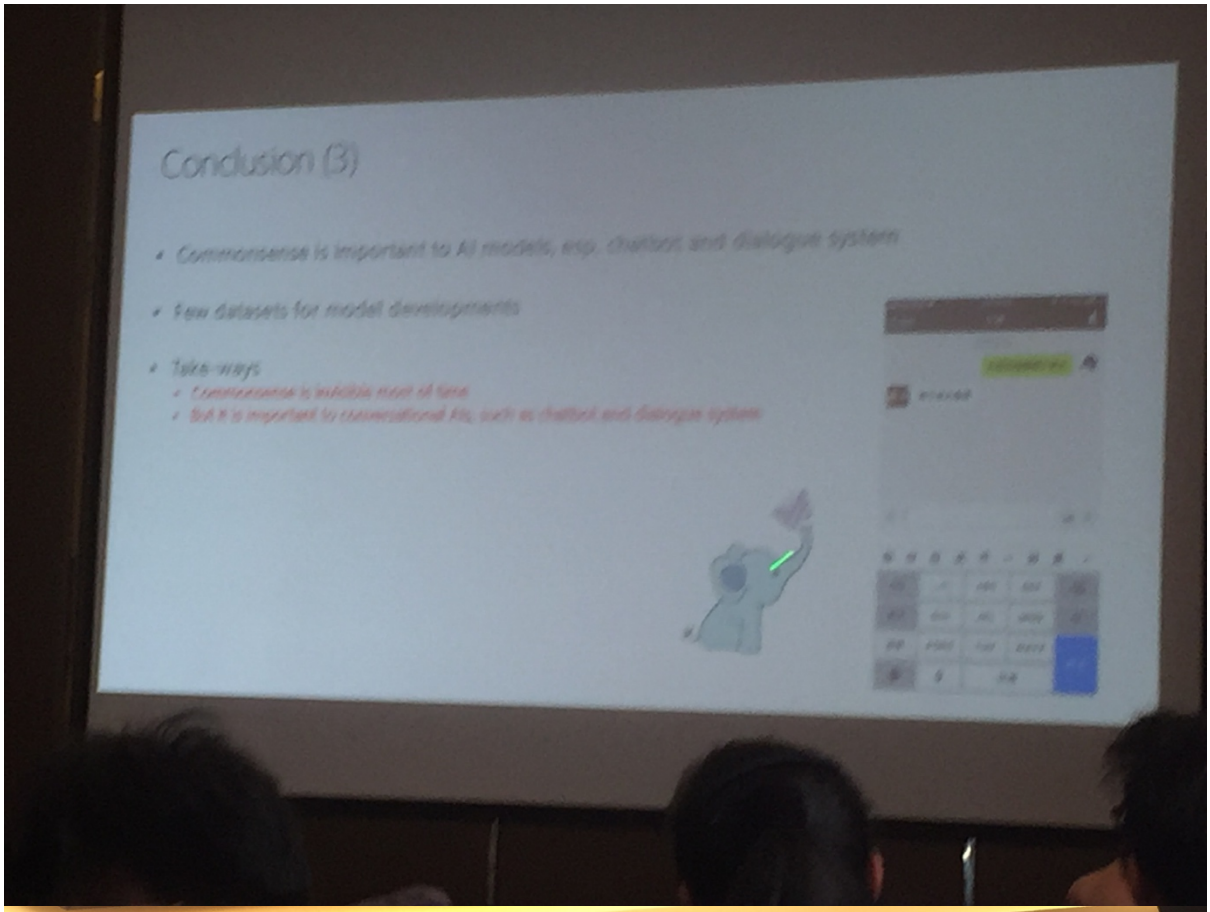
Winograd Schema Challenge
The trophy doesn't fit in the suitcase because it is too big.
Question: What is too big?
Answer: (a) the trophy (b) the suitcase



Monolingual Corpus



Language Model for Commonsense Reasoning
(Trinh and Le, 2018)

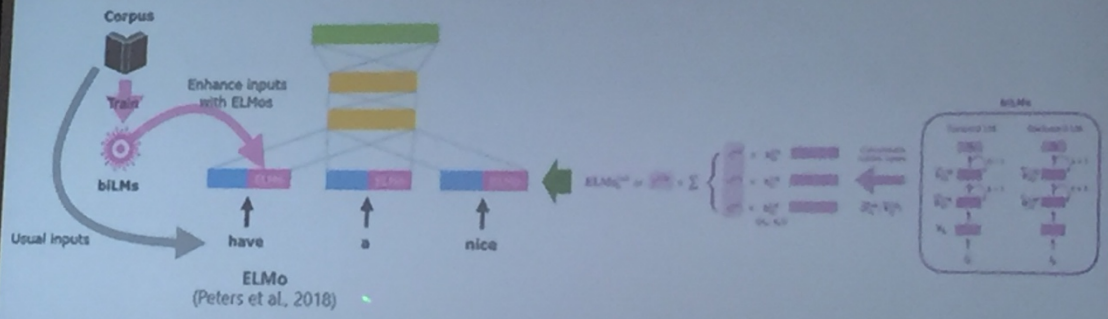

Conclusion (3)
Commonsense is important to AI models, esp. chatbot and dialogue system
Few datasets for model developments
Take-ways


Pre-trained Task Embeddings
Encoder-Decoder framework is widely used in NLP tasks
Encoders of different tasks have a similar structure
Each encoder can be considered as the task knowledge representation
Bill Gates is the founder of Microsoft
LM Encoder
Monolingual Data
Bill Gates is the founder of ___
MT Encoder
Bilingual Data
Paraphrase Encoder
Paraphrase Data
Microsoft was founded by whom

Transfer Learning with Pre-trained Task Embeddings in NLP

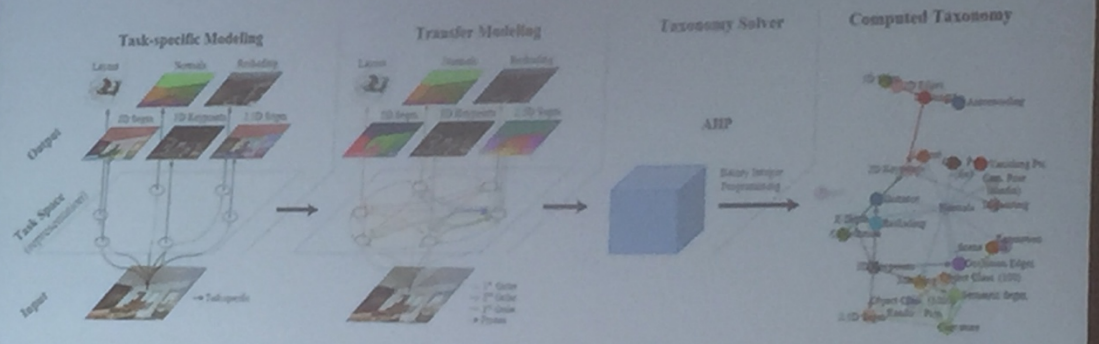
- Knowledge acquired by a trained model can be **transferred** during training process for a new task

ELMo can be integrated to almost all neural NLP tasks
with simple concatenation to the embedding layer



ELMo
(Peters et al., 2018)

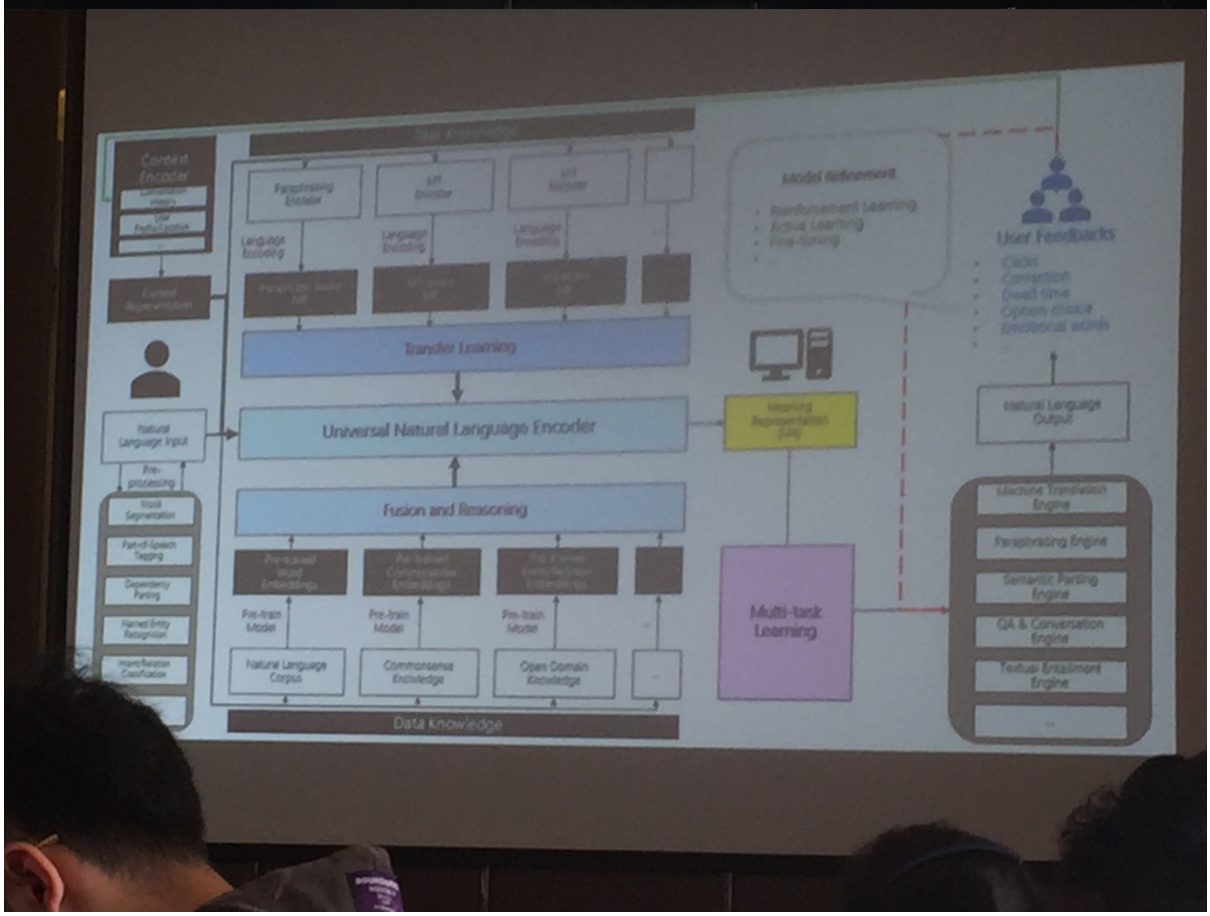
Transfer Learning with Pre-trained Task Embeddings in CV



(Zamir et al., 2018): The task taxonomy (**Taskonomy**) is a computationally found directed hypergraph that captures the notion of task transferability over any given task dictionary.

Conclusion (4)

- Take-aways
  - Transfer learning is important to NLP



Summary and Challenge

- Models can learn **implicit knowledge** from large-scale datasets for rich-resource tasks!
- Models need **explicit knowledge** for low-resource tasks without much training data!
- How to **define**, **acquire** and **represent** knowledge?
- How to **represent meanings of language** with knowledge?
- How to **reason** with knowledge?
- How to **transfer knowledge** from one task/modal to another task/modal?
- How to **memorize**, **update** and **forget** knowledge as context?

段楠, 周明. 《智能问答》. 高等教育出版社, 2018 (in press)