THE NATURAL LANGUAGE DECATHLON: MULTITASK LEARNING AS QUESTION ANSWERING

Bryan McCann, Nitish Shirish Keskar,
Caiming Xiong, Richard Socher
Salesforce Research

Presenter: Happy Buzaaba

October 6th 2019

MLT Paper Reading Session

Agenda

- Introduction
- Motivation
- Technical Contribution
- Approach
- Evaluation & Results
- Conclusion

Introduction

Focus in this paper is multi-task learning. 10 different NLP tasks are jointly learned.

Natural language decathlon (decaNLP) is the benchmark for measuring the performance of the NLP models across ten tasks that appear disparate until unified as question answering.

Introduction

- Deep learning has improved performance on many NLP tasks individually.

  - **Single-task learning**: Great performance improvement in recent years given {dataset, task, model and metric}

    As long as data is plentiful, we can hill-climb to local optima

Introduction

- Deep learning has improved performance on many NLP tasks individually.

  - Limits of Single-task learning

    1. New task with dataset and metric
    2. New Model
    3. Train (almost) from scratch
    4. Repeat

human like

- However, general NLP models cannot emerge within a paradigm that focuses on the particularities of a single metric, dataset, and task.

What is Next for Natural Language Processing?

- Multi-task learning:

1. Question Answering
2. Machine Translation
3. Summarization
4. Natural language Inference
5. Sentiment analysis
6. Semantic role labeling
7. Relation extraction
8. Goal oriented dialogue
9. Semantic parsing
10. Commonsense pronoun resolution

- Think of which ones might form a basis set of tasks that would help the model understand many different features of language and allow them to design a model that is not particular to any task but can solve all the tasks they want it to work on.

Motivation

- Why a single Multi-task learning model:

1. Step towards general AI/NLP models and Ideas
2. Easy to adopt new tasks
3. Easier deployment in production
4. Lowering the bar for any body to solve their NLP tasks

Main Contribution

- Introduce the Natural Language Decathlon (decaNLP), a challenge that spans 10 tasks

- Cast all tasks as question answering over a context.

- Present a new multitask question answering network (MQAN) that jointly learns all tasks in decaNLP without any task-specific modules or parameters more effectively than seq2seq and reading comprehension baselines.

Natural Language Decathlon (decaNLP) Overview

Examples

Question	Context	Answer
What is a major importance of Southern California in relation to California and the US?	...Southern California is a major economic center for the state of California and the US....	major economic center
What is the translation from English to German?	Most of the planet is ocean water.	Der Großteil der Erde ist Meerwasser
What is the summary?	Harry Potter star Daniel Radcliffe gains access to a reported £320 million fortune...	Harry Potter star Daniel Radcliffe gets £320M fortune...
Hypothesis: Product and geography are what make cream skimming work. Entailment, neutral, or contradiction?	Premise: Conceptually cream skimming has two basic dimensions – product and geography.	Entailment
Is this sentence positive or negative?	A stirring, funny and finally transporting re-imagining of Beauty and the Beast and 1930s horror film.	positive

Question	Context	Answer
What has something experienced?	Areas of the Baltic that have experienced eutrophication.	eutrophication
Who is the illustrator of Cycle of the Werewolf?	Cycle of the Werewolf is a short novel by Stephen King, featuring illustrations by comic book artist Bernie Wrightson.	Bernie Wrightson
What is the change in dialogue state?	Are there any Eritrean restaurants in town?	food: Eritrean
What is the translation from English to SQL?	The table has column names... Tell me what the notes are for South Australia	SELECT notes from table WHERE 'Current Slogan' = 'South Australia'
Who had given help? Susan or Joan?	Joan made sure to thank Susan for all the help she had given.	Susan

- Dataset with one example from each decaNLP task. Each task is framed as a form of QA. Answer words in red, generated by pointing to the context, in green from the question and in blue if they are generated from the classifier over the full output vocabulary.

Approach
Task definition: Multitask Learning as Question Answering

- Meta-Supervised Learning: From {x, y} to {x, t, y} (t is the task)[1]
- Use a question, q, as a natural description of the task, t, to allow the model to use linguistic information to connect tasks
- y is the answer to q and x is the context necessary to answer q

Note: Meta learning includes the task **t** in training set to orient the model to the correct task. Which allows single models to effectively multi-task, it makes them more suitable as pretrained models, and allows the model to generalize to completely new tasks through different but related contexts and questions.

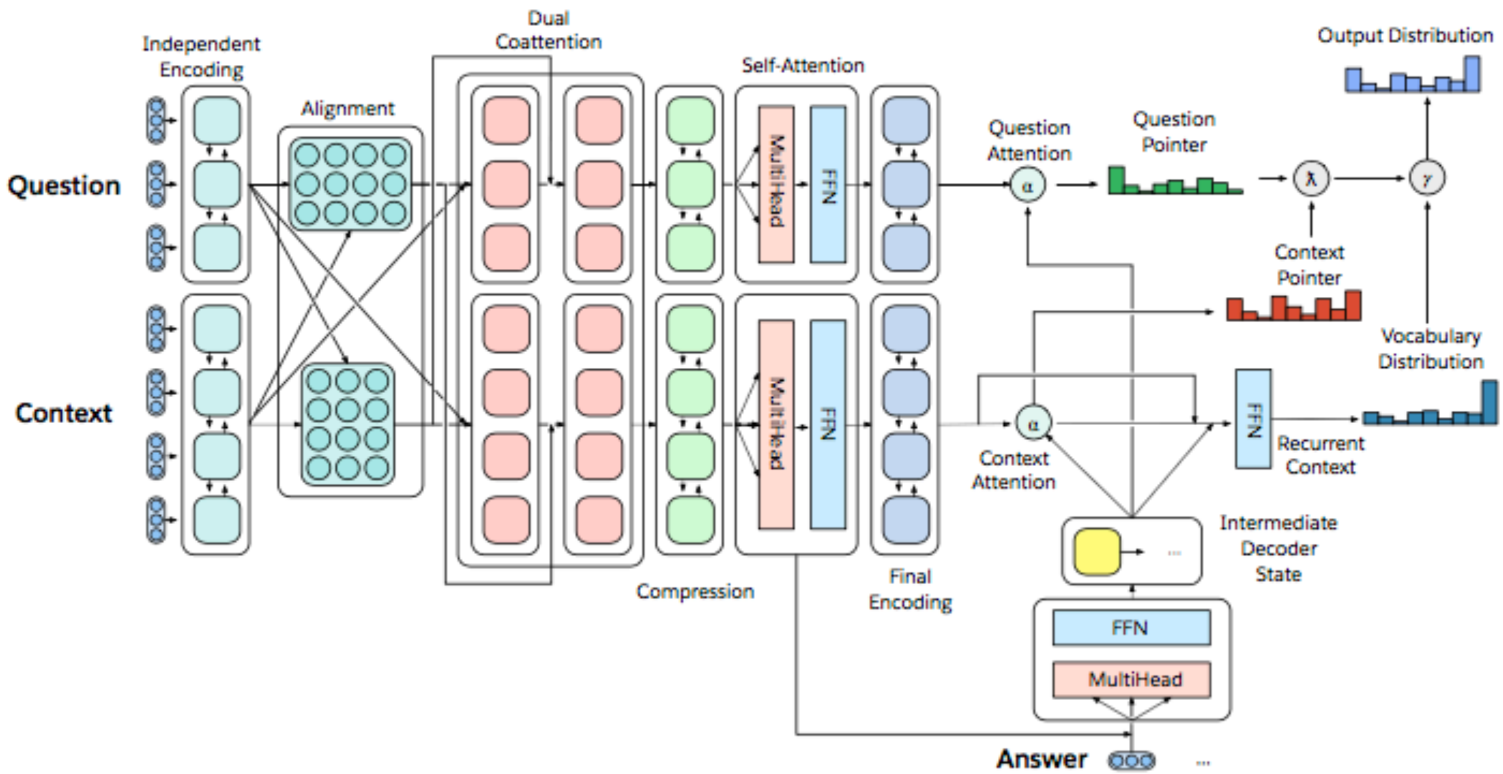
[1][Schmidhuber, 1987, Thrun and Pratt, 1998, Thrun, 1998, Vilalta and Drissi, 2002]

Task definition: Designing a model for decaNLP

Specifications:

- No task-specific modules or parameters because we assume the task ID is not available
- Must be able to adjust internally to perform disparate tasks
- Should leave open the possibility to zero-shot inference for un seen tasks

Multi-task Question Answering Network (MQAN) for decaNLP)



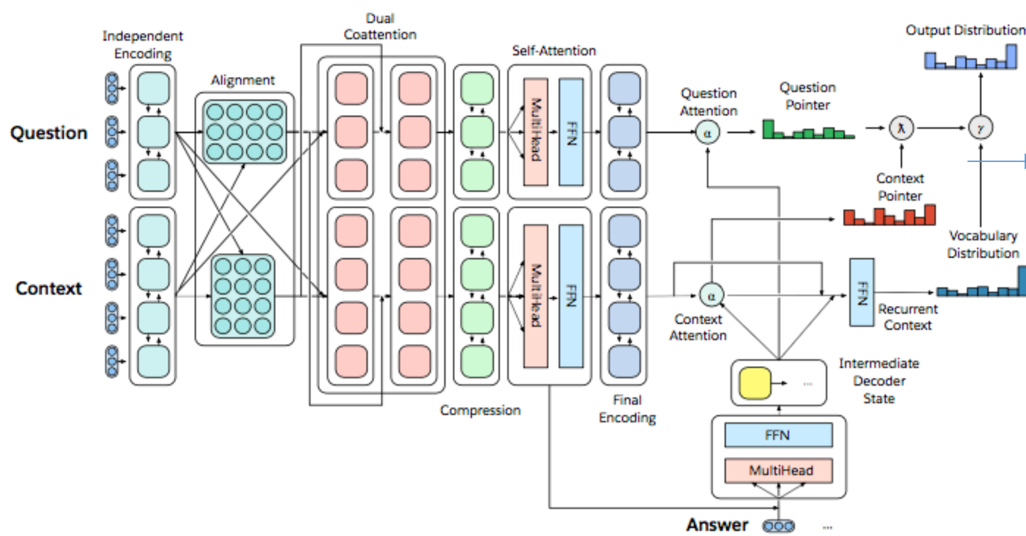
Multi-task Question Answering Network (MQAN) for decaNLP)

Context
the **table** has column names… tell me what the notes are for South Australia

Question
what is the translation from English to SQL?

Answer
select notes from table where current slogan

- Start with context
- Ask a question
- Generate answer one word at a time by
  - pointing to context
  - pointing to question
  - Or choosing word from external vocabulary



Pointer switch

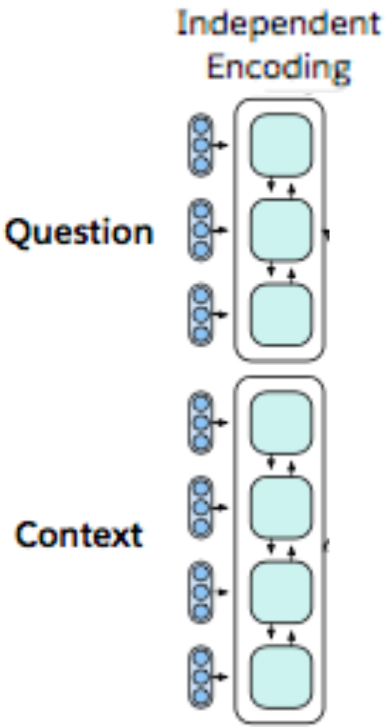
Question pointer

Context pointer

Vocabulary

Note: Pointer switch is choosing from the three options for each output

Multi-task Question Answering Network (MQAN) for decaNLP)



Input matrices

$$Q \in R^{m \times demb}$$

$$C \in R^{l \times demb}$$

Projected to a common $\boldsymbol{d}$ (linear layer)

$$QW_1 = Q_{proj} \in R^{m \times d}$$

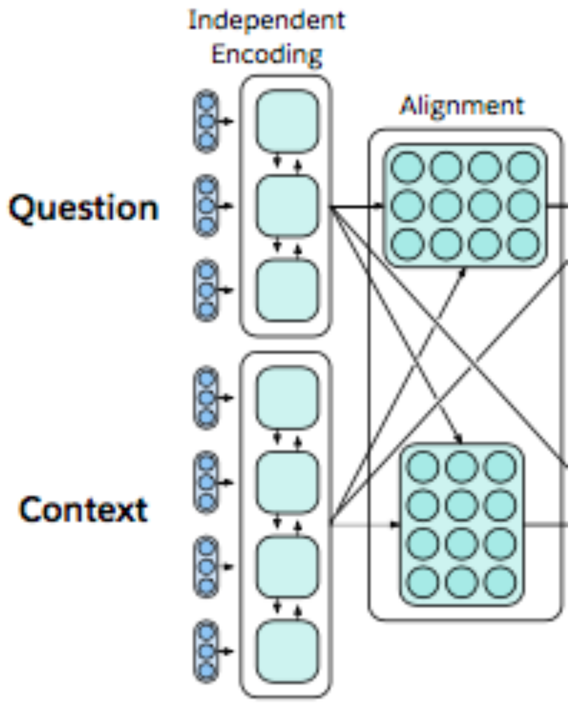
$$CW_1 = C_{proj} \in R^{l \times d}$$

Final BiLSTM layer for both question and context

$$\mathbf{BiLSTM}_{ind}(Q_{proj}) = Q_{ind} \in \mathrm{R}^{m \times d} \qquad \mathbf{BiLSTM}_{ind}(C_{proj}) = C_{ind} \in \mathrm{R}^{l \times d}$$

Fixed Glove+Character n-gram embeddings → linear → Shared BiLSTM with skip connection

Multi-task Question Answering Network (MQAN) for decaNLP)



Add trained dummy embeddings

$C_{ind} \in \mathrm{R}^{(l+1)\times d}$

$Q_{ind} \in \mathrm{R}^{(m+1)\times d}$

Avoid forcing tokens
To align with any token
in other sequence

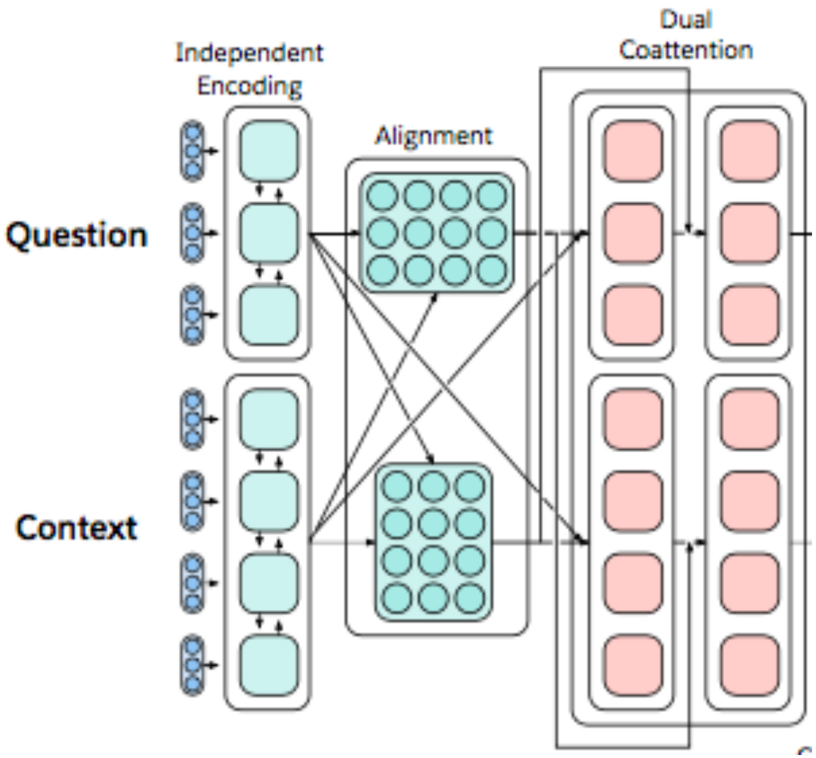
Alignments

$$softmax\,(Q_{ind}C^{\top}_{ind}) = S_{qc} \in R^{(m+1)\times(l+1)}$$

$$softmax\,(C_{ind}Q^{\top}_{ind}) = S_{cq} \in \mathrm{R}^{(l+1)\times(m+1)}$$

We obtain coattended representations by first aligning encoded representations of each sequence

Multi-task Question Answering Network (MQAN) for decaNLP)



Weighted summations

$$S^{\top}_{qc}Q_{ind} = Q_{sum} \in R^{(l+1)\times d}$$

$$S^{\top}_{cq}C_{ind} = C_{sum} \in R^{(m+1)\times d}$$

Transfer alignment info to original sequence

$$S^{\top}_{qc}C_{sum} = C_{coa} \in R^{(l+1)\times d}$$

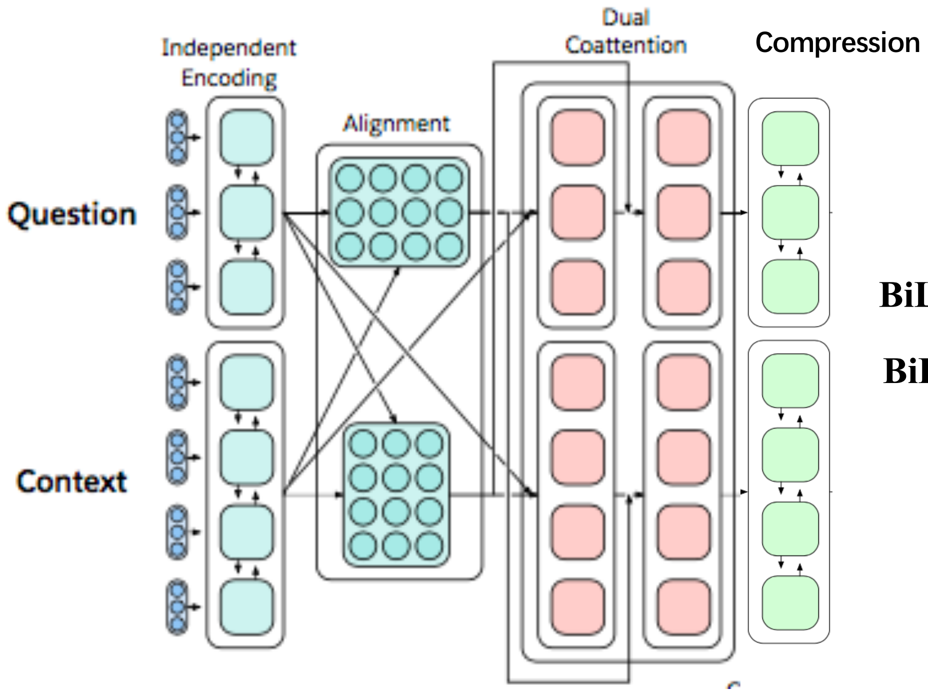
$$S^{\top}_{cq}Q_{sum} = Q_{coa} \in R^{(m+1)xd}$$

Drop dummy embedding

$$Q_{coa} \in R^{mxd} \qquad C_{coa} \in R^{l\times d}$$

Attention summations from one sequence to the other and back again with skip connections

Multi-task Question Answering Network (MQAN) for decaNLP)



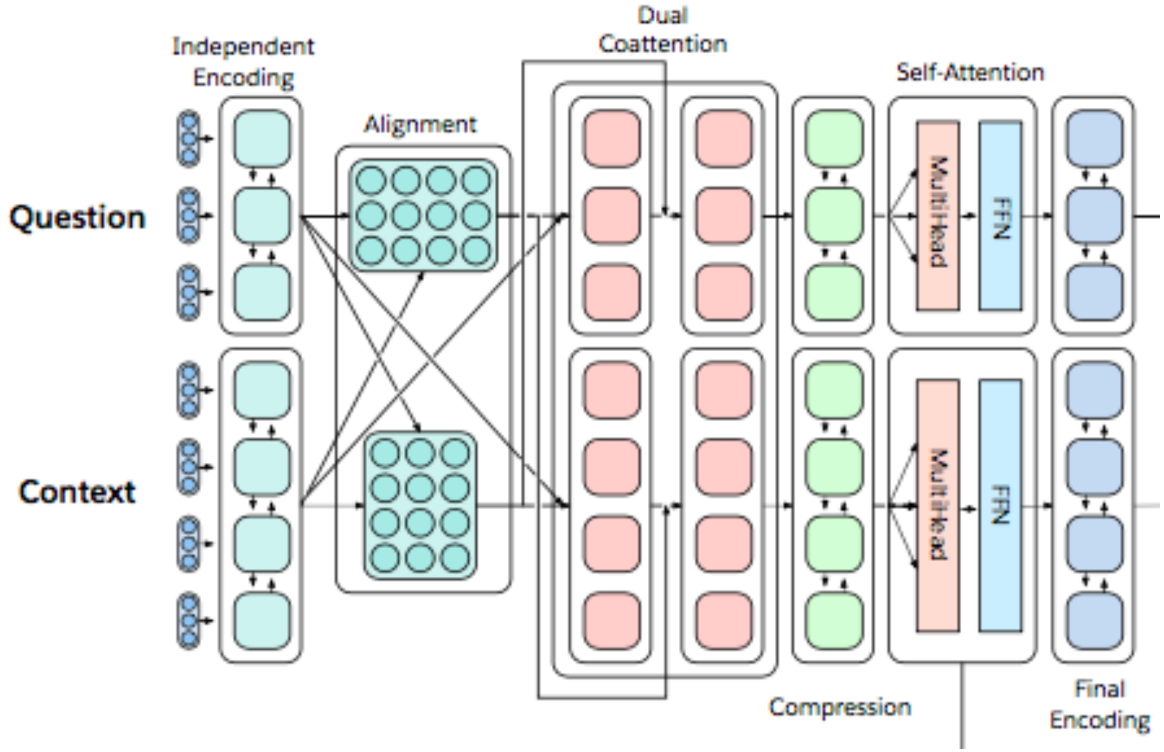
Concatenate all seq representations

$$\mathbf{BiLSTM}_{com}\boldsymbol{Q}\ ([\boldsymbol{Q}_{proj}\ ;\ \boldsymbol{Q}_{ind};\ \boldsymbol{C}_{sum};\ \boldsymbol{Q}_{coa}]) = \boldsymbol{Q}_{com} \in \boldsymbol{R}^{lxd}$$

$$\mathbf{BiLSTM}_{com}\boldsymbol{C}\ ([\boldsymbol{C}_{proj}\ ;\ \boldsymbol{C}_{ind};\ \boldsymbol{Q}_{sum};\ \boldsymbol{C}_{coa}]) = \boldsymbol{C}_{com} \in \boldsymbol{R}^{lxd}$$

we concatenate all four prior representations for each sequence along the last dimension and feed into separate BiLSTM

Multi-task Question Answering Network (MQAN) for decaNLP)



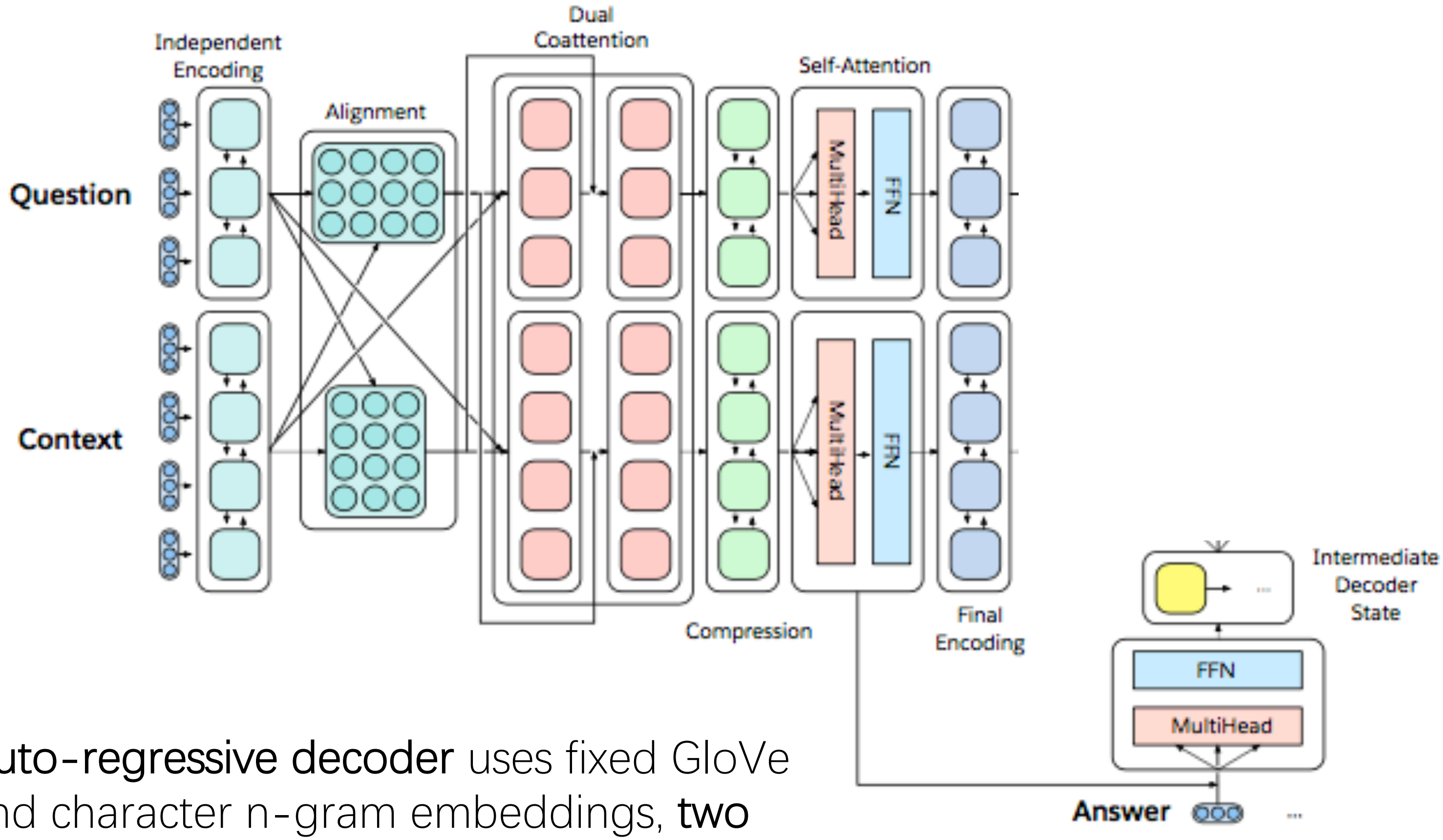
$$\mathbf{BiLSTM}_{finQ}(Q_{self}) = Q_{fin} \in \mathrm{R}^{m\times d}$$

$$\mathbf{BiLSTM}_{finC}(C_{self}) = C_{fin} \in \mathrm{R}^{l\times d}$$

Separate BiLSTMS to reduce dimensionality, two transformer layers, another BiLSTM

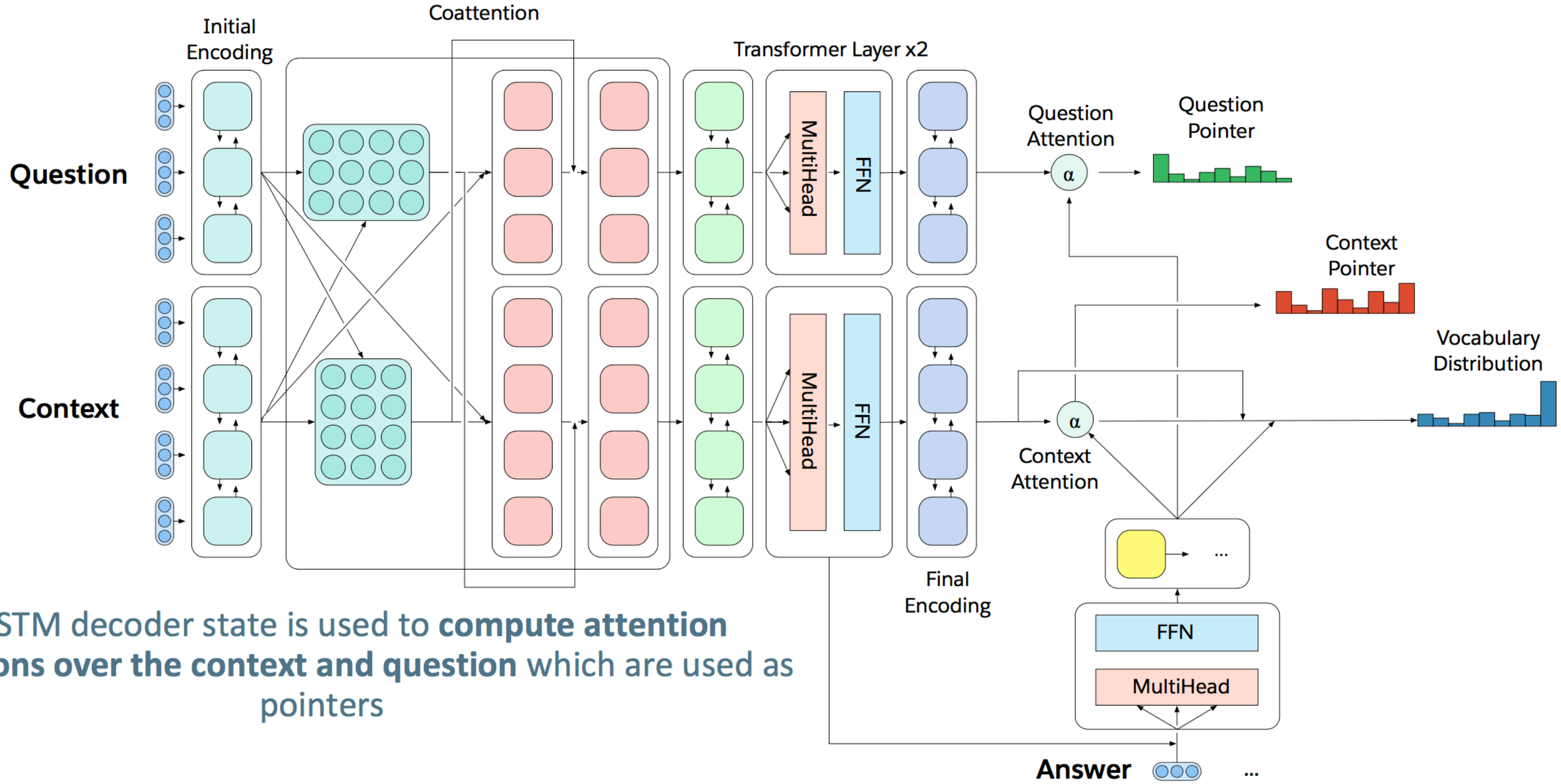
Attention is all you need, NIPS 2017

Multi-task Question Answering Network (MQAN) for decaNLP)



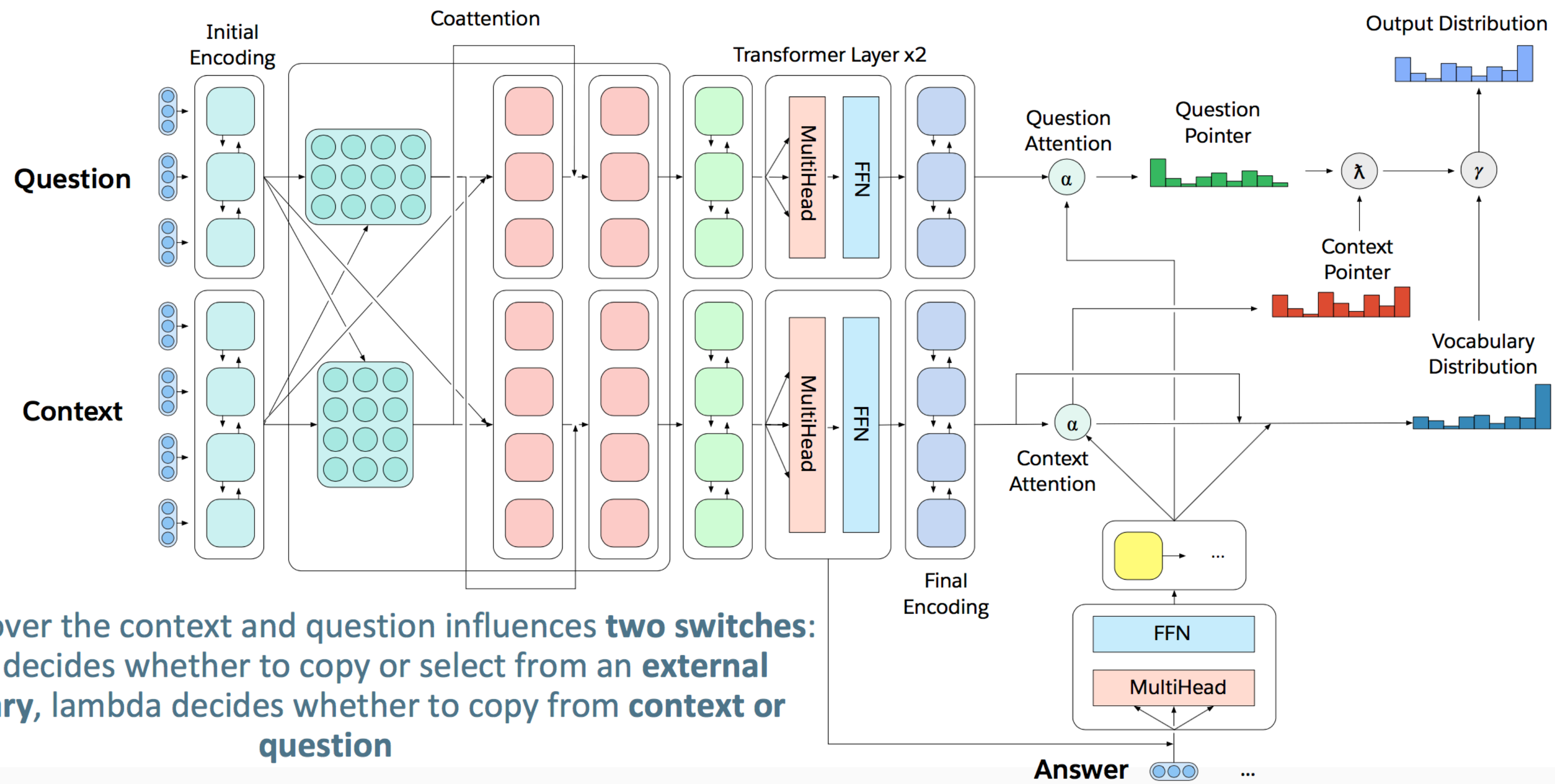
Auto-regressive decoder uses fixed GloVe and character n-gram embeddings, **two transformer layers** and an **LSTM layer** that attend to outputs of the last three layers of the encoder

Multi-task Question Answering Network (MQAN) for decaNLP)



The LSTM decoder state is used to **compute attention distributions over the context and question** which are used as pointers

Multi-task Question Answering Network (MQAN) for decaNLP)



Attention over the context and question influences **two switches**: gamma decides whether to copy or select from an **external vocabulary**, lambda decides whether to copy from **context or question**

Evaluation

Datasets: Different metrics for different tasks

Task	Dataset	# Train	# Dev	# Test	Metric
Question Answering	SQuAD	87599	10570	9616	nF1
Machine Translation	IWSLT	196884	993	1305	BLEU
Summarization	CNN/DM	287227	13368	11490	ROUGE
Natural Language Inference	MNLI	392702	20000	20000	EM
Sentiment Analysis	SST	6920	872	1821	EM
Semantic Role Labeling	QA-SRL	6414	2183	2201	nF1
Zero-Shot Relation Extraction	QA-ZRE	840000	600	12000	cF1
Goal-Oriented Dialogue	WOZ	2536	830	1646	dsEM
Semantic Parsing	WikiSQL	56355	8421	15878	lfEM
Pronoun Resolution	MWSC	80	82	100	EM

nF1: normalized F1 that strips out articles and punctuation
ROUGE: Average of ROUGE-1, 2, and L
EM : exact match comparison (text classification: accuracy)
dsEM: turn-based dialogue state exact match
lfEM: logical forms exact match
cF1: corpus level metric (takes into account that some questions are unanswerable)

Evaluation

Datasets: Different metrics for different tasks

Task	Dataset	# Train	# Dev	# Test	Metric
Question Answering	SQuAD	87599	10570	9616	nF1
Machine Translation	IWSLT	196884	993	1305	BLEU
Summarization	CNN/DM	287227	13368	11490	ROUGE
Natural Language Inference	MNLI	392702	20000	20000	EM
Sentiment Analysis	SST	6920	872	1821	EM
Semantic Role Labeling	QA-SRL	6414	2183	2201	nF1
Zero-Shot Relation Extraction	QA-ZRE	840000	600	12000	cF1
Goal-Oriented Dialogue	WOZ	2536	830	1646	dsEM
Semantic Parsing	WikiSQL	56355	8421	15878	lfEM
Pronoun Resolution	MWSC	80	82	100	EM

Natural Language Decathlon decascore

decascore = sum of task-specific metrics

Results

Datasets: Different metrics for different tasks

	Single-task Training				Multitask Training			
Dataset				**(+QPtr)**				**(+QPtr)**
SQuAD				75.3				70.8
IWSLT				26.7				16.1
CNN/DM				25.5				23.9
MNLI				73				70.5
SST				88.5				86.2
QA-SRL				77.9				75.8
QA-ZRE				24.3				28
WOZ				88				80.6
WikiSQL				73.5				62
MWSC				48.8				48.8
decaScore	518.8	559.2	537.2	601.5	**513.6**	**546.4**	**533.8**	**562.7**

S2S: is the sequence to sequence baseline
(+Satt): S2S + self Attention in encoder
(+Catt): S2S + self Attention and coattention in the encoder

Results

Datasets: Different metrics for different tasks

	Single-task Training				Multitask Training			
Dataset	**S2S**	**(+Satt)**	**(+Catt)**	**(+QPtr)**	**S2S**	**(+Satt)**	**(+Catt)**	**(+QPtr)**
SQuAD	48.2	68.2	74.6	75.3	47.5	66.8	71.8	70.8
IWSLT	25	23.3	26	26.7	14.2	13.6	9	16.1
CNN/DM	19	20	25.1	25.5	25.7	14	15.7	23.9
MNLI	67.5	68.5	34.7	73	60.9	69	70.4	70.5
SST	86.4	86.8	86.2	88.5	85.9	84.7	86.5	86.2
QA-SRL	63.5	67.8	74.8	77.9	68.7	75.1	76.1	75.8
QA-ZRE	20	19.9	16.6	24.3	28.5	31.7	28.5	28
WOZ	85.3	86	86.5	88	84	82.8	75.1	80.6
WikiSQL	60	72.4	72.3	73.5	45.8	64.8	62.9	62
MWSC	43.9	46.3	40.4	48.8	52.4	43.9	37.8	48.8
decaScore	518.8	559.2	537.2	601.5	**513.6**	**546.4**	**533.8**	**562.7**

Transformer layers, yield benefits in single-task and multitask setting

Results

Datasets: Different metrics for different tasks

	Single-task Training				Multitask Training			
Dataset	**S2S**	**(+Satt)**	**(+Catt)**	**(+QPtr)**	**S2S**	**(+Satt)**	**(+Catt)**	**(+QPtr)**
SQuAD	48.2	68.2	74.6	75.3	47.5	66.8	71.8	70.8
IWSLT	25	23.3	26	26.7	14.2	13.6	9	16.1
CNN/DM	19	20	25.1	25.5	25.7	14	15.7	23.9
MNLI	67.5	68.5	34.7	73	60.9	69	70.4	70.5
SST	86.4	86.8	86.2	88.5	85.9	84.7	86.5	86.2
QA-SRL	63.5	67.8	74.8	77.9	68.7	75.1	76.1	75.8
QA-ZRE	20	19.9	16.6	24.3	28.5	31.7	28.5	28
WOZ	85.3	86	86.5	88	84	82.8	75.1	80.6
WikiSQL	60	72.4	72.3	73.5	45.8	64.8	62.9	62
MWSC	43.9	46.3	40.4	48.8	52.4	43.9	37.8	48.8
decaScore	518.8	559.2	537.2	601.5	**513.6**	**546.4**	**533.8**	**562.7**

Transformer layers, yield benefits in single-task and multitask setting
Question answering and semantic role labeling have a strong connection.

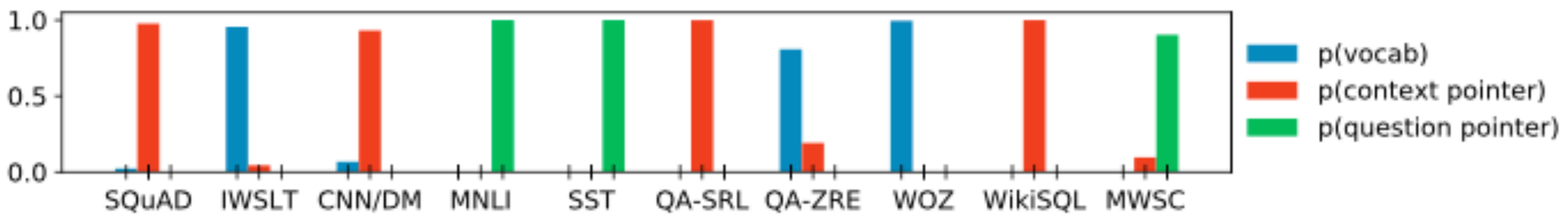
Results

Datasets: Different metrics for different tasks

	Single-task Training				Multitask Training			
Dataset	**S2S**	**(+Satt)**	**(+Catt)**	**(+QPtr)**	**S2S**	**(+Satt)**	**(+Catt)**	**(+QPtr)**
SQuAD	48.2	68.2	74.6	75.3	47.5	66.8	71.8	70.8
IWSLT	25	23.3	26	26.7	14.2	13.6	9	16.1
CNN/DM	19	20	25.1	25.5	25.7	14	15.7	23.9
MNLI	67.5	68.5	34.7	73	60.9	69	70.4	70.5
SST	86.4	86.8	86.2	88.5	85.9	84.7	86.5	86.2
QA-SRL	63.5	67.8	74.8	77.9	68.7	75.1	76.1	75.8
QA-ZRE	20	19.9	16.6	24.3	28.5	31.7	28.5	28
WOZ	85.3	86	86.5	88	84	82.8	75.1	80.6
WikiSQL	60	72.4	72.3	73.5	45.8	64.8	62.9	62
MWSC	43.9	46.3	40.4	48.8	52.4	43.9	37.8	48.8
decaScore	518.8	559.2	537.2	601.5	**513.6**	**546.4**	**533.8**	**562.7**

Transformer layers, yield benefits in single-task and multitask setting
Question answering and semantic role labeling have a strong connection.
There is a gap between the combined single task models and the single multitask model

Evaluation



- Answers are correctly copied from either context or question

- No confusion over which tasks the model should perform or which output space to use

Evaluation

Pretraining on decaNLP improves final performance

- An example: Additional IWSLT Language pairs (Left)
- New tasks like NER (Right)

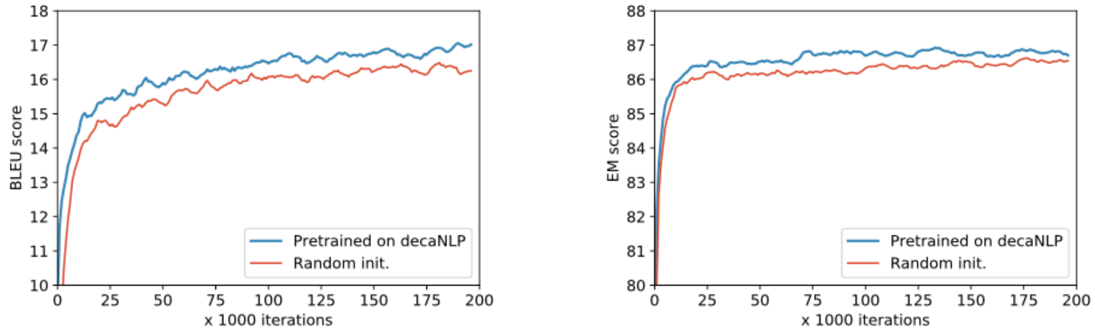


Figure 4: MQAN pretrained on decaNLP outperforms random initialization when adapting to new domains and learning new tasks. Left: training on a new language pair – English to Czech, right: training on a new task – Named Entity Recognition (NER).

Conclusion

- We introduce (decaNLP) a new benchmark for measuring the performance of NLP models across 10 tasks that appear disparate until unified as QA.
- We present MQAN a model for general question answering that uses muti-pointer-generator decoder to capitalize on questions as natural language descriptions of tasks
- MQAN exhibits transfer-learning and zero-short capabilities.