What do you learn from context?

Probing for sentence structure in contextualized word representations.

Review by Gong Qiong

Outline

- Abstract
- Introduction
- Edge Probing
- Experiment Design
- Results and Analysis
- Discussion

Abstract

- Contextualized representation models such as ELMo and BERT have recently achieved state-of-the-art results on a diverse array of downstream NLP tasks.
- Building on recent token-level probing work, we introduce a novel edge probing task design and construct a broad suite of sub-sentence tasks derived from the traditional structured NLP pipeline.

Abstract

- We probe word-level contextual representations from four recent models and investigate how they encode sentence structure across a range of syntactic, semantic, local, and long-range phenomena.
- We find that existing models trained on language modeling and translation produce strong representations for syntactic phenomena, but only offer comparably small improvements on semantic tasks over a non-contextual baseline.

Introduction

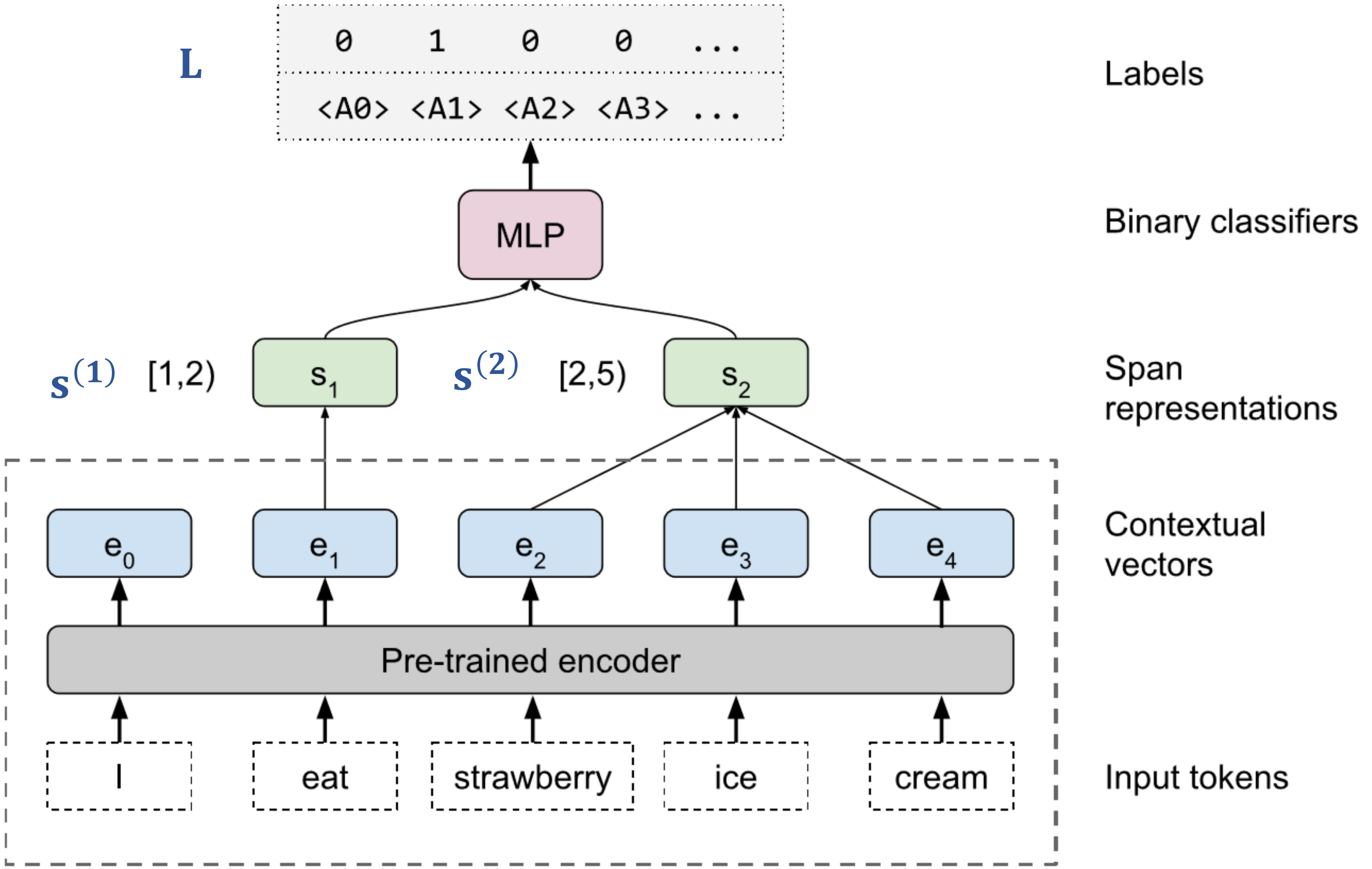
Introduction - 2 Qs

What do you learn from context

- **Syntactic** or **Semantic**?
- **Local** or **Long-range**?

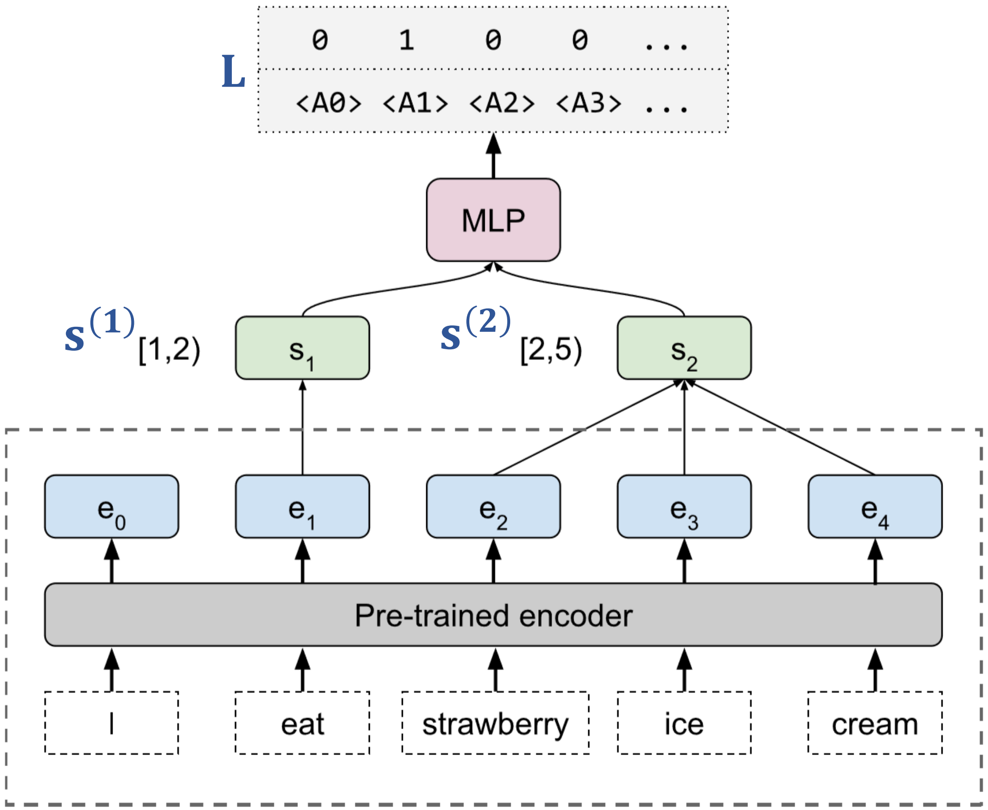
What do you learn from context?

Edge probing - model architecture



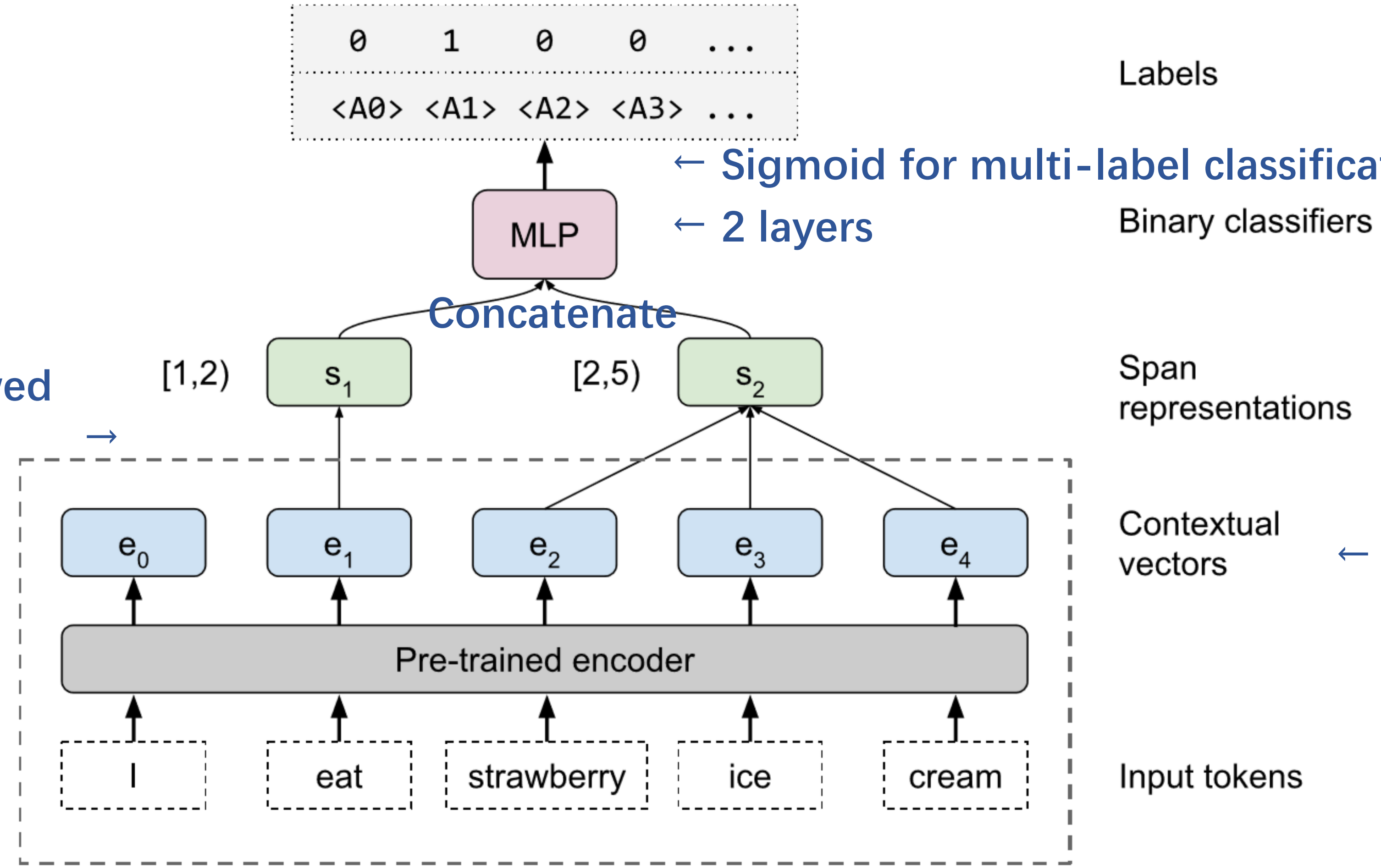
Edge probing

- Represent a sentence as a list of tokens
  - $T = [t_0, t_1, \dots, t_n]$
- Input a labeled edge as $\{s^{(1)}, s^{(2)}, L\}$
  - $s^{(1)} = [i^{(1)}, j^{(1)})$
  - $s^{(2)} = [i^{(2)}, j^{(2)})$, optional (as Coreference), omitted for unary edges (as POS tagging)
  - L: a set of one or more targets from a task-specific label set.
- Predict L as multi-label classification
  - **Uniform metric**: binary F1 score



Common architecture

Edge probing - model architecture



← Sigmoid for multi-label classification

← 2 layers

← Fixed-length

← Fixed & Pre-trained

Projection layer followed by the self-attention pooling operator →

Edge probing

- **Separate projections**
  - Why? -> Different regulations can be extracted from different parts ($s^{(1)}, s^{(2)}$)
- **Self-Attention Pooling**
  - 1. Calculate a weight.
  - 2. Compute a weighted span representations.
  - Why? -> Strengthen the model so that it can perform better, making the result more obvious and prominent.
- **MLP: 2 layers**

Experiment Design

NLP Tasks

- Syntactic tasks
  - Part-of-speech tagging (POS)
  - Constituent labeling
  - Dependency labeling
  - Named entity labeling
  - Semantic role labeling (SRL) core roles
- Semantic tasks
  - Semantic proto-role (SPR)
  - Semantic role labeling (SRL) non-core roles
  - Coreference (OntoNote + Winograd-style)
  - Relation Classification (Rel.)

Various!

NLP Tasks - example

0 1 2 3 4 5 6 7 8 9 10 11

POS	The important thing about Disney is that it is a global [brand]$_1$. $\rightarrow$ NN (Noun)
Constit.	The important thing about Disney is that it [is a global brand]$_1$. $\rightarrow$ VP (Verb Phrase)
Depend.	[Atmosphere]$_1$ is always [fun]$_2$ $\rightarrow$ nsubj (nominal subject)
Entities	The important thing about [Disney]$_1$ is that it is a global brand. $\rightarrow$ Organization
SRL	[The important thing about Disney]$_2$ [is]$_1$ that it is a global brand. $\rightarrow$ Arg1 (Agent)
SPR	[It]$_1$ [endorsed]$_2$ the White House strategy... $\rightarrow$ {awareness, existed_after, ...}
Coref.[O]	The important thing about [Disney]$_1$ is that [it]$_2$ is a global brand. $\rightarrow$ True
Coref.[W]	[Characters]$_2$ entertain audiences because [they]$_1$ want people to be happy. $\rightarrow$ True Characters entertain [audiences]$_2$ because [they]$_1$ want people to be happy. $\rightarrow$ False
Rel.	The [burst]$_1$ has been caused by water hammer [pressure]$_2$. $\rightarrow$ Cause-Effect(e_2, e_1)

Table 1: Example sentence, spans, and target label for each task. O = OntoNotes, W = Winograd.

Contextualized Representation Models

Model	Training objective/task	Structure	Training Corpus	Usage
CoVe	Machine Translation	Two-layer biLSTM	WMT2017: 7 million sentences from web crawl, news, and government proceedings	Top-level activation, concatenated with Glove vectors. (300*2+300 dims)
ELMo	Language Modeling	Two-layer bidirectional LSTM over a context-independent character CNN layer	Billion Word Benchmark	Standard usage
GPT		12-layer Transformer as a left-to-right language model	Toronto Books Corpus	Do not fine-tune, **cat(as CoVe) /mix(as ELMo)**
BERT	Masked Language Modeling (MLM) & Next Sentence Prediction (NSP)	Deep Transformer (12/24-layer)	Toronto Books Corpus & English Wikipedia	

Baselines

- **Lexical Baselines**
  - Contextualized vs Non-contextualized
  - CoVe vs 300 dims GloVe
  - ELMo vs context-independent character-CNN of ELMo
  - GPT&BERT vs subword embeddings

- **Randomized ELMo**
  - Random orthonormal matrices: Invalidate the information of ELMo.

- **Word-Level CNN**
  - CNN: See tokens around the center word.

Results and Analysis

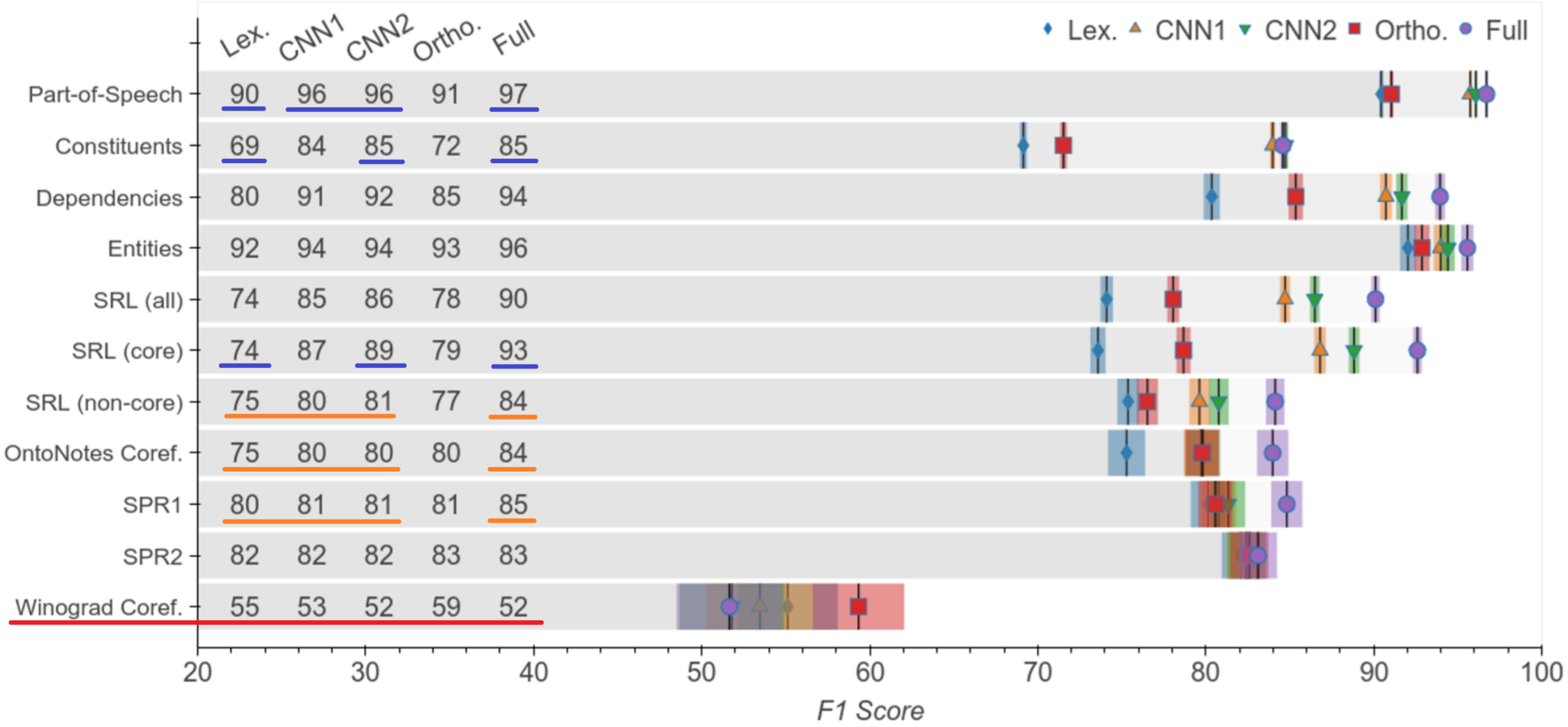
Results and Analysis

- CoVe, ELMo, GPT, BERT
  - the large gains: mostly syntactic
  - the small gains: mostly semantic
- SRL (core roles) > SRL (non-core roles)
- mix > cat
- Large improvement of BERT on Winograd-style coreference
- GPT trained on BWB ≈ GPT trained on Books Corpus

	CoVe			ELMo			GPT		
	Lex.	Full	Abs. Δ	Lex.	Full	Abs. Δ	Lex.	cat	mix
Part-of-Speech	85.7	94.0	8.4	90.4	**96.7**	6.3	88.2	94.9	95.0
Constituents	56.1	81.6	25.4	69.1	**84.6**	15.4	65.1	81.3	**84.6**
Dependencies	75.0	83.6	8.6	80.4	**93.9**	13.6	77.7	92.1	**94.1**
Entities	88.4	90.3	1.9	92.0	**95.6**	3.5	88.6	92.9	92.5
SRL (all)	59.7	80.4	20.7	74.1	**90.1**	16.0	67.7	86.0	89.7
Core roles	*56.2*	*81.0*	*24.7*	*73.6*	***92.6***	*19.0*	*65.1*	*88.0*	*92.0*
Non-core roles	*67.7*	*78.8*	*11.1*	*75.4*	***84.1***	*8.8*	*73.9*	*81.3*	***84.1***
OntoNotes coref.	72.9	79.2	6.3	75.3	84.0	8.7	71.8	83.6	**86.3**
SPR1	73.7	77.1	3.4	80.1	**84.8**	4.7	79.2	83.5	83.1
SPR2	76.6	80.2	3.6	82.1	83.1	1.0	82.2	**83.8**	**83.5**
Winograd coref.	52.1	**54.3**	2.2	**54.3**	**53.5**	-0.8	51.7	52.6	**53.8**
Rel. (SemEval)	51.0	60.6	9.6	55.7	77.8	22.1	58.2	**81.3**	**81.0**
Macro Average	69.1	78.1	9.0	75.4	**84.4**	9.1	73.0	83.2	**84.4**

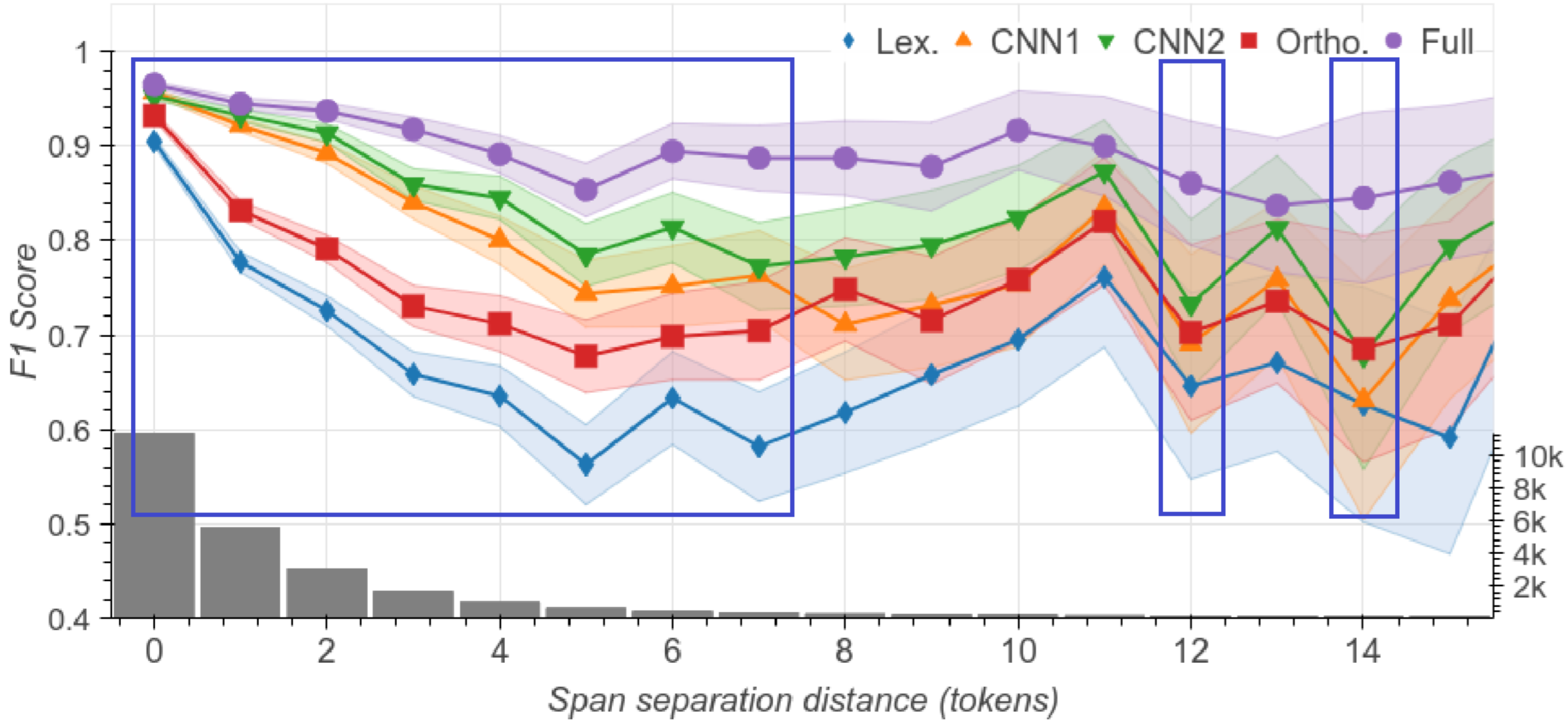
	BERT-base				BERT-large				
	F1 Score			Abs. Δ	F1 Score			Abs. Δ	
	Lex.	cat	mix	ELMo	Lex.	cat	mix	(base)	ELMo
Part-of-Speech	88.4	**97.0**	96.7	0.0	88.1	96.5	**96.9**	0.2	0.2
Constituents	68.4	83.7	86.7	2.1	69.0	80.1	**87.0**	0.4	2.5
Dependencies	80.1	93.0	95.1	1.1	80.2	91.5	**95.4**	0.3	1.4
Entities	90.9	96.1	96.2	0.6	91.8	96.2	**96.5**	0.3	0.9
SRL (all)	75.4	89.4	91.3	1.2	76.5	88.2	**92.3**	1.0	2.2
Core roles	*74.9*	*91.4*	*93.6*	*1.0*	*76.3*	*89.9*	***94.6***	*1.0*	*2.0*
Non-core roles	*76.4*	*84.7*	*85.9*	*1.8*	*76.9*	*84.1*	***86.9***	*1.0*	*2.8*
OntoNotes coref.	74.9	88.7	90.2	6.3	75.7	89.6	**91.4**	1.2	7.4
SPR1	79.2	84.7	**86.1**	1.3	79.6	85.1	**85.8**	-0.3	1.0
SPR2	81.7	83.0	**83.8**	0.7	81.6	83.2	**84.1**	0.3	1.0
Winograd coref.	54.3	53.6	54.9	1.4	53.0	53.8	**61.4**	6.5	7.8
Rel. (SemEval)	57.4	78.3	82.0	4.2	56.2	77.6	**82.4**	0.5	4.6
Macro Average	75.1	84.8	86.3	1.9	75.2	84.2	**87.3**	1.0	2.9

Results and Analysis



- **Word-Level CNN closes most of the gap on syntactic tasks.**
  - **Local information**
- **The gap between Word-Level and full model is large still on semantic tasks.**
  - **Long-range information**
- Full model > Orthonormal encoder > Lexical baseline
  - **A flaw: Why so weird on Winograd coref.?**

Results and Analysis



- Full ELMo holds up better, dropping only 7 F1 score from d=0 to d=8
  - The pre-trained encoder does encode useful long-distance dependencies.

Results and Analysis - summary

- **Syntactic** or **Semantic**? -> **Syntactic** > **Semantic**

- **Local** or **Long-range**? -> **Local** & **Long-range**

- Different layer, different info.
- Well-designed architecture
- Deep model does well on difficult semantic task

Discussion

What can we learn from that? & What should we do in the future?

Discussion

- Existing representation models encode more syntactic than semantic info.
  - Bright future: To capture semantic context information

- BERT(-base) outperforms GPT, average 1.6-1.9 F1 score
  - Novel effective training objectives
  - Baidu-ERNIE: ‘xxx is one of the top universities in the world.’

- No common sense -> Using knowledgebase
  - Learn semantics and common sense

References

- Tenney I, Xia P, Chen B, et al. What do you learn from context? Probing for sentence structure in contextualized word representations[J]. 2018.
- McCann B, Bradbury J, Xiong C, et al. Learned in translation: Contextualized word vectors[C]//Advances in Neural Information Processing Systems. 2017: 6294-6305.
- Peters M E, Neumann M, Iyyer M, et al. Deep contextualized word representations[J]. arXiv preprint arXiv:1802.05365, 2018.
- Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training[J]. URL https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf , 2018.
- BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding