Have We Solved The Hard Problem? It’s Not Easy! Contextual Lexical Contrast as a Means to Probe Neural Coherence

AAAI 2021
Full Slides

Wenqiang Lei, Yisong Miao, Runpeng Xie, Bonnie Webber, Meichun Liu, Tat-Seng Chua and Nancy F. Chen

https://cont2lex.github.io/

Table of Contents

1. Task Introduction — Contextual Lexical Contrast

Contextual Lexical Contrast (CLC)

Example: positive vs negative:

(Ex. 1 Positive CLC): A **positive** attitude helps you relax and ace the exams, and a **negative** mental status will however make you nervous and sleepless.

(Ex. 2 Negative CLC): The reviewers are rather **positive** about this paper. They are nominating it for the Best Paper for its discovery of a **negative** finding that dispels conventional wisdom.

Definition of CLC (a new NLP task):

- Two words from the same sentence (or adjacent sentences) form a "contextual lexical contrast" word pair if these two words exhibit contrastive semantics that contribute to the coherence of sentential context.

2. Motivation and Background

Motivation and Background

— Why CLC is important.

- Cohesion Modeling
  - Entity-based
  - Lexical-based
- Lexical Contrast and Lexical Relation
- Interpretations of Semantic Representations

Cohesion Modeling

— Lexical-based approach is overlooked.

Entity-based Approach

BERT (Bidirectional Encoder Representations from Transformers; Devlin et al., 2019) is also based on the transformer, but it is bidirectional as opposed to left-to-right as in the OpenAI GPT, and the directions are dependent as opposed to ELMo's independently trained left-to-right and right-to-left LSTMs. It also introduces a somewhat different objective called ''masked language model'': during training, some tokens are randomly masked, and the objective is to restore them from the context.

Excerpted from Shwartz & Dagan, TACL2019

✅ Entity grid method (Barzilay and Lapata)

Lexical-based Approach

BERT (Bidirectional Encoder Representations from Transformers; Devlin et al., 2019) is also based on the transformer, but it is bidirectional as opposed to left-to-right as in the OpenAI GPT, and the directions are dependent as opposed to ELMo's independently trained left-to-right and right-to-left LSTMs. It also introduces a somewhat different objective called ''masked language model'': during training, some tokens are randomly masked, and the objective is to restore them from the context.

Excerpted from Shwartz & Dagan, TACL2019

❓

- Being Largely Ignored
- Need to put into context

Lexical Contrast
— Context is critical for downstream applications.

Computing Lexical Contrast

Saif M. Mohammad*
National Research Council Canada

Bonnie J. Dorr**
University of Maryland

Graeme Hirst†
University of Toronto

Peter D. Turney‡
National Research Council Canada

Computational Linguistic, 2013

Applications

- Discourse relation.

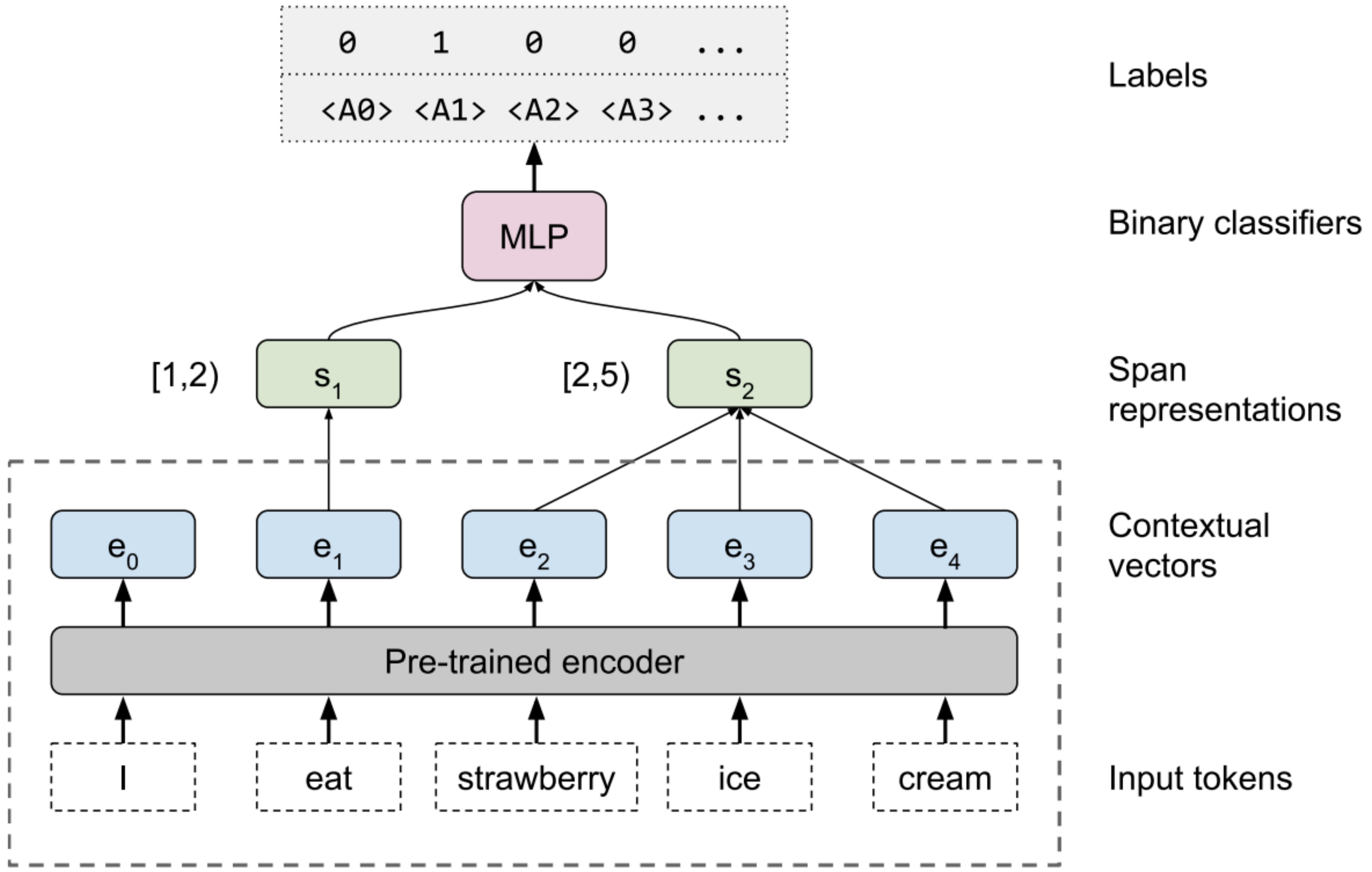
"Tokyo is ***cold***. Beijing is ***hot***."

- Contradiction detection.

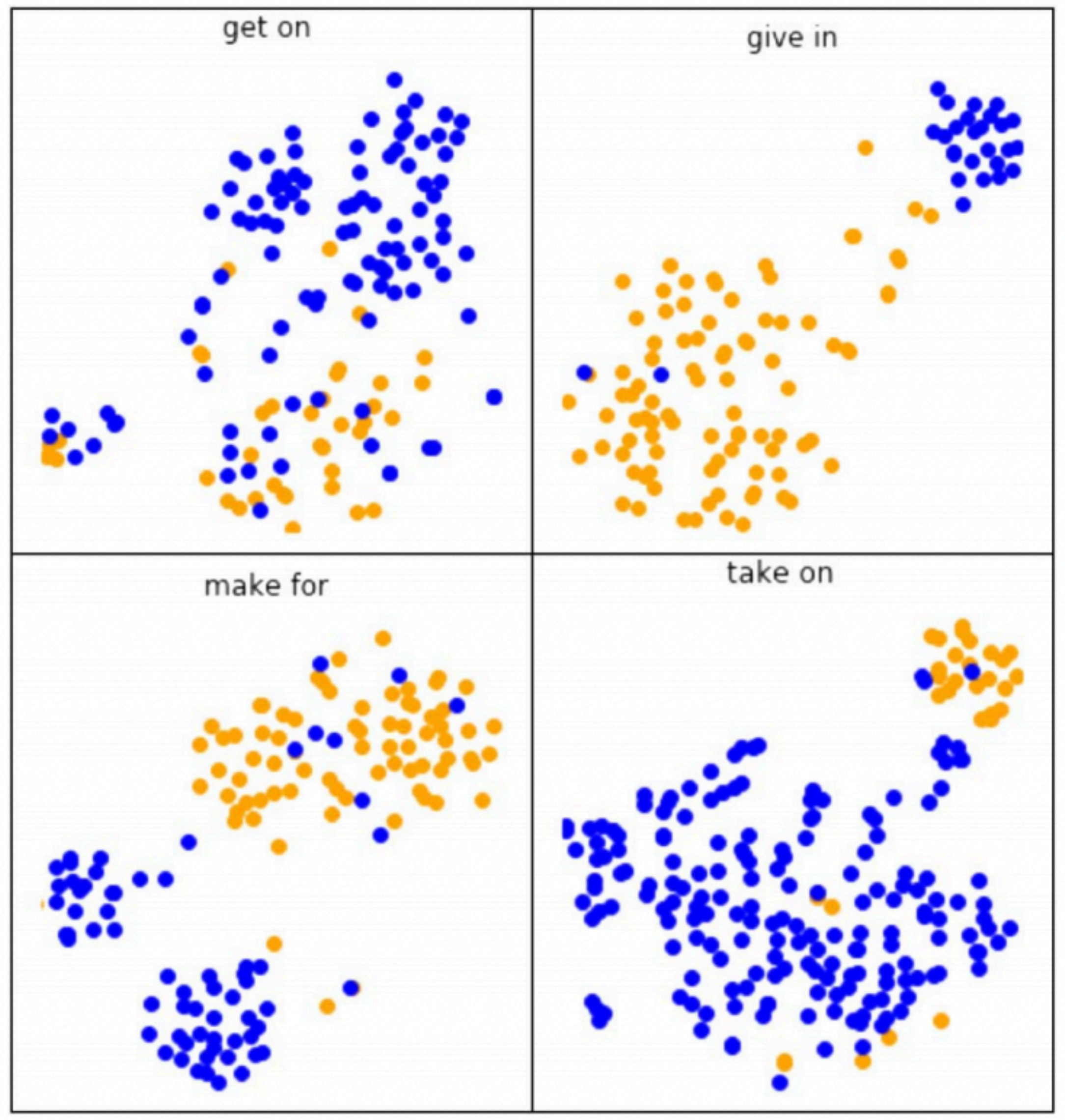
"Kyoto has a predominantly ***wet*** climate" / "It is mostly ***dry*** in Kyoto"

- Humour detection.

Interpretations of Semantic Representations
— Right timing to do CLC.



Probing Contextual LMs (Tenney et.al. ICLR '19)



Probing Contextual Lexical Composition (Shwartz and Dagan TACL '19)

- Syntactic tasks: POS, Constituents, Dependencies
- Semantic tasks: SRL, OntoNotes coref, Semantic proto-role

- Light Verb Construction (LVC): *make* a decision
- Verb-Particle Construction (VPC): carry on vs carry

3. Cont^{2}Lex Corpus

Problem Formalization

Problem Formalization:

Given w^{+} and w^{-} in context c (a sequence of words $w_1, w_2, \ldots w_n$), a human (or a machine) needs to indicate a binary tag for CLC.

Instance Preparation

ConceptNet
An open, multilingual knowledge graph

Wikipedia
The Free Encyclopedia

WSJ

- Constraint 1: Contrasting degree in ConceptNet
- Constraint 2: Distance between w^+ and w^- (Adjacent sentence or difference clause in same sentence.)
- Constraint 3: Appearance of the same pair of w^+ and w^-

6,316 instances to be annotated.

Human Annotation

(Ex. 1 Positive CLC): A <u>**positive**</u> attitude helps you relax and ace the exams, and a <u>**negative**</u> mental status will however make you nervous and sleepless.

(Ex. 2 Negative CLC): The reviewers are rather <u>**positive**</u> about this paper. They are nominating it for the Best Paper for its discovery of a <u>**negative**</u> finding that dispels conventional wisdom.

- Quality Control 1: Predict w^-, given only w^+ and c
- Quality Control 2: Hard-to-decide Option.

Corpus Statistics

Inter-Annotator Agreement (IAA):
We calculate IAA using the consensus of our 5 annotators, reaching 75.3%.

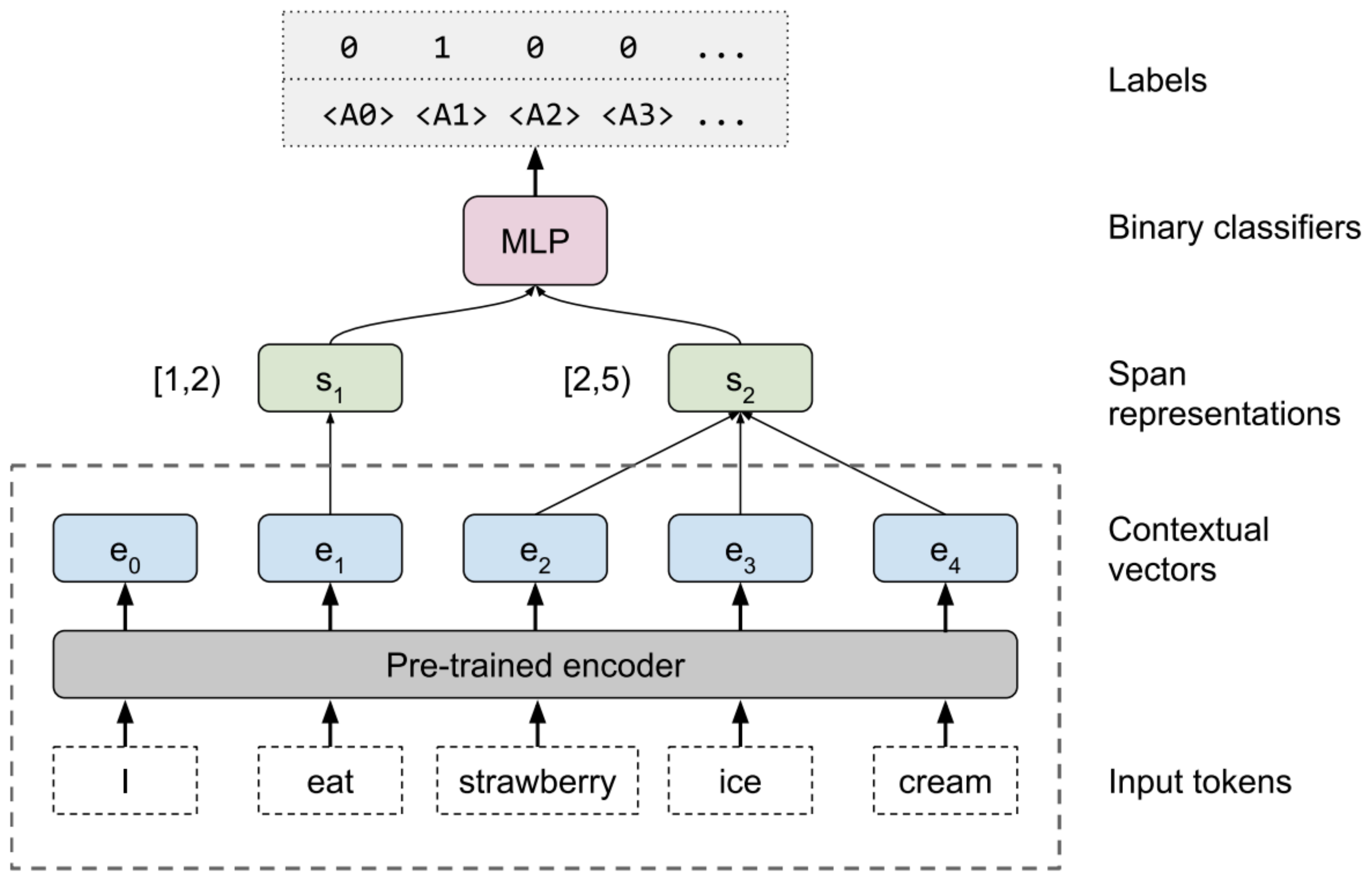
Part-of-Speech	#	Positive Ratio
Noun	2,413	33.2%
Verb	1,568	27.9%
Adj	2,081	43.7%
Adv	254	40.9%
Total	6,316	35.7%

Possible reason: Adj and Adv has purer semantic dimensions.

4. Benchmark

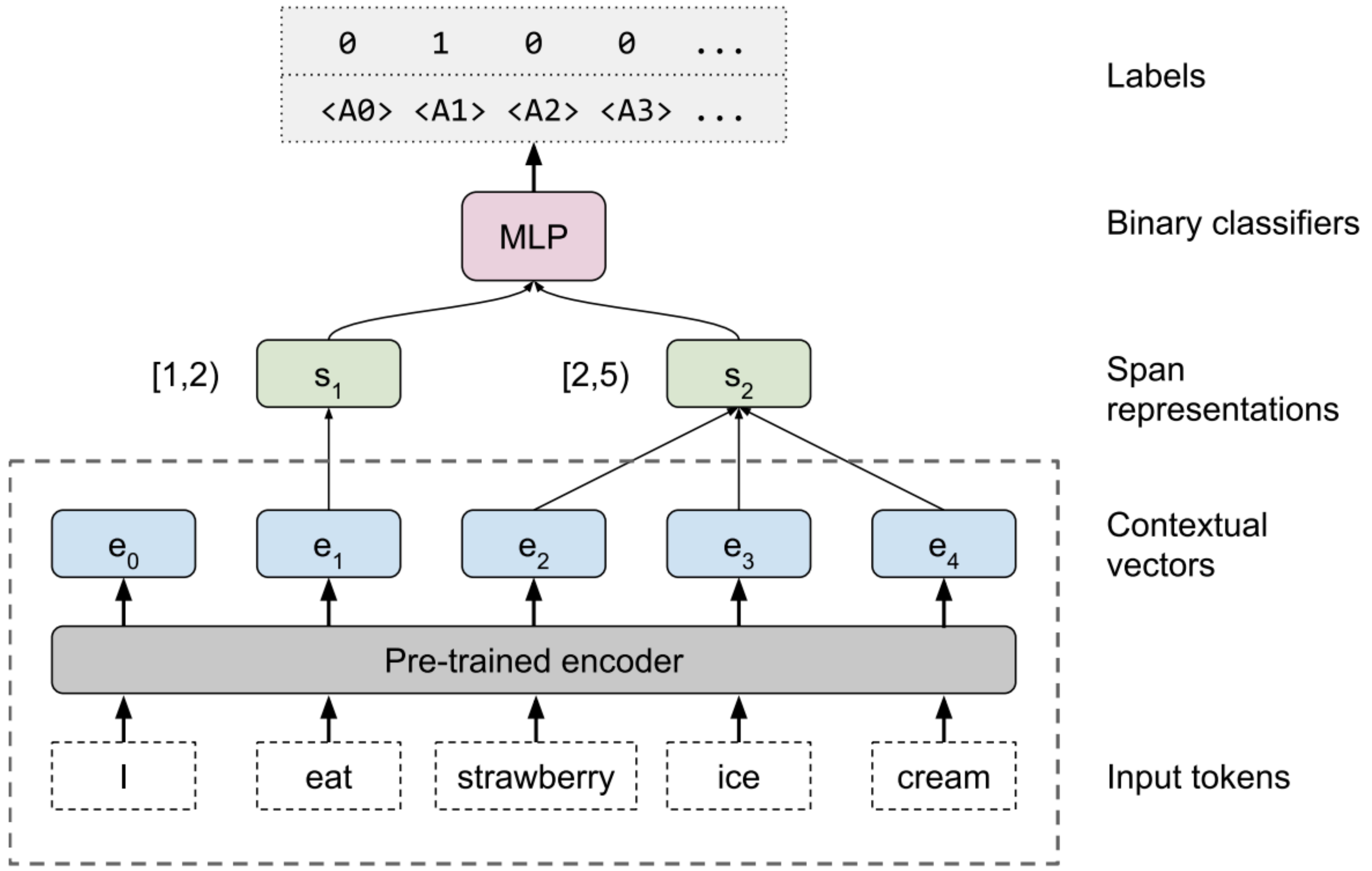
Evaluation Framework

- 6,316 instances enable us to do supervised learning, for the binary classification.
- Similar approach as Tenney et.al, and “Embed — Encode — Predict” framework (Shwartz and Dagan)
- We didn’t fine-tune BERT. Why?



Probing Contextual LMs (Tenney et.al. ICLR ’19)

Evaluated Embeddings



Probing Contextual LMs (Tenney et.al. ICLR ’19)

- Static embeddings: Glove, Word2Vec, fastText
- Contextual Embeddings: ELMo, OpenAI GPT, BERT
- The “Lex” version of GPT and BERT. Why?

5. Experiments and Conclusion

Research Questions

- RQ1: How do models perform on the CLC recognition?
- RQ2: Are models able to recognize lexical contrast out-of-context?
- RQ3: What are the capabilities and limitations of current models?

Main Experiment (RQ1)

	BiLSTM	Attention	None
Glove	65.3	64.9	65.3
Word2Vec	65	65.7	64.7
FastText	66.2	65.5	66.3
ELMo	65.6	65.6	65.7
GPT.Lex	65.8	64.8	64.8
GPT	66.8	67.0	66.9
BERT.Lex	66.4	66.2	66.4
BERT	70.0	69.2	69.1
Majority	64.3		

BERT and GPT are better than their Lex version.

Acc scores show that CLC is a challenging task!

Out-of-context Lexical Contrast (RQ2)

(Ex. 2 Negative CLC): The reviewers are rather **positive** about this paper. They are nominating it for the Best Paper for its discovery of a **negative** finding that dispels ***conventional*** wisdom.

👆

Embeddings	Glove	Word2Vec	fastText	ELMo	GPT	BERT
acc.	79.7	82.6	84.1	83.5	81.2	79.5

Acc scores of out-of-context lexical contrast recognition, which is much more higher than CLC.

[Contextual]$_{\text{hard}}$ Lexical [Contrast]$_{\text{easy}}$

Model Characteristics (RQ3)

S: CLC Word Pairs Occurring in the Same Sentence.

R: Word Repetitions Co-Occurring with CLC Pairs.

(Ex. 3 Repetition): ...is considered <u>spurious</u> **by** Hefele questionable by Haddan and Stubbs, and <u>genuine</u> **by** JaffA Regest.

(Ex. 4 Repetition): They had many children who **lived in the** <u>darkness</u> between them. The children wished to **live in the** <u>light</u> and so separated their unwilling parents.

Model Characteristics (RQ3)

	S	¬S	R	¬R
Glove+None	61.3 (+4.2)	67.9 (-2.0)	60.9 (+7.2)	67.3 (-3.1)
W2V+Attention	60.3 (+3.2)	68.8 (-1.1)	60.4 (+6.7)	68.1 (-2.3)
FastText+None	60.4 (+3.3)	69.8 (-0.1)	61.1 (+7.4)	68.8 (-1.6)
ELMo+None	63.6 (+6.5)	68 (-1.9)	63 (+9.4)	68 (-2.5)
GPT.Lex+BiLSTM	61.5 (+4.4)	68.3 (-1.6)	60.8 (+7.1)	68.1 (-2.3)
GPT+Attention	64 (+6.9)	68.7 (-1.2)	65.5 (+11.8)	67.8 (-2.6)
BERT.Lex+BiLSTM	60.7 (+3.6)	69.8 (-0.1)	58.7 (+5.0)	69.9 (-0.4)
BERT+BiLSTM	67.4 (+10.3)	71.4 (+1.5)	68.7 (+14.9)	70.7 (+0.3)
Majority	57.1	69.9	53.7	70.4

The delta over baseline are majorly achieved by S and R.

Model Characteristics (RQ3)

— Q: Besides Repetition, what other cohesive ties is BERT using?

Cohesive devices (M.A.K. Halliday):

- Collocation
- Substitution
- Coreference

T: All types of cohesive ties

R: Repetition

R is a subset of T.

	¬R	¬T
ΔBERT+BiLSTM	4.1	4.2
ΔBERT+Attention	3.6	3.5
ΔBERT+None	3.7	3.7

This table shows that models are no better handling T than R.

Conclusion

- We propose a new NLP task as CLC for cohesion modelling. Our Cont^{2}Lex corpus makes CLC a computational feasible task.

- CLC is a challenging semantic representation task. Contextual embeddings are capable to capture part of contextual information.

- The advantage gained by BERT is largely due to modelling surface textual patterns.