Understanding Idiomatic Langauge using Neural Networks

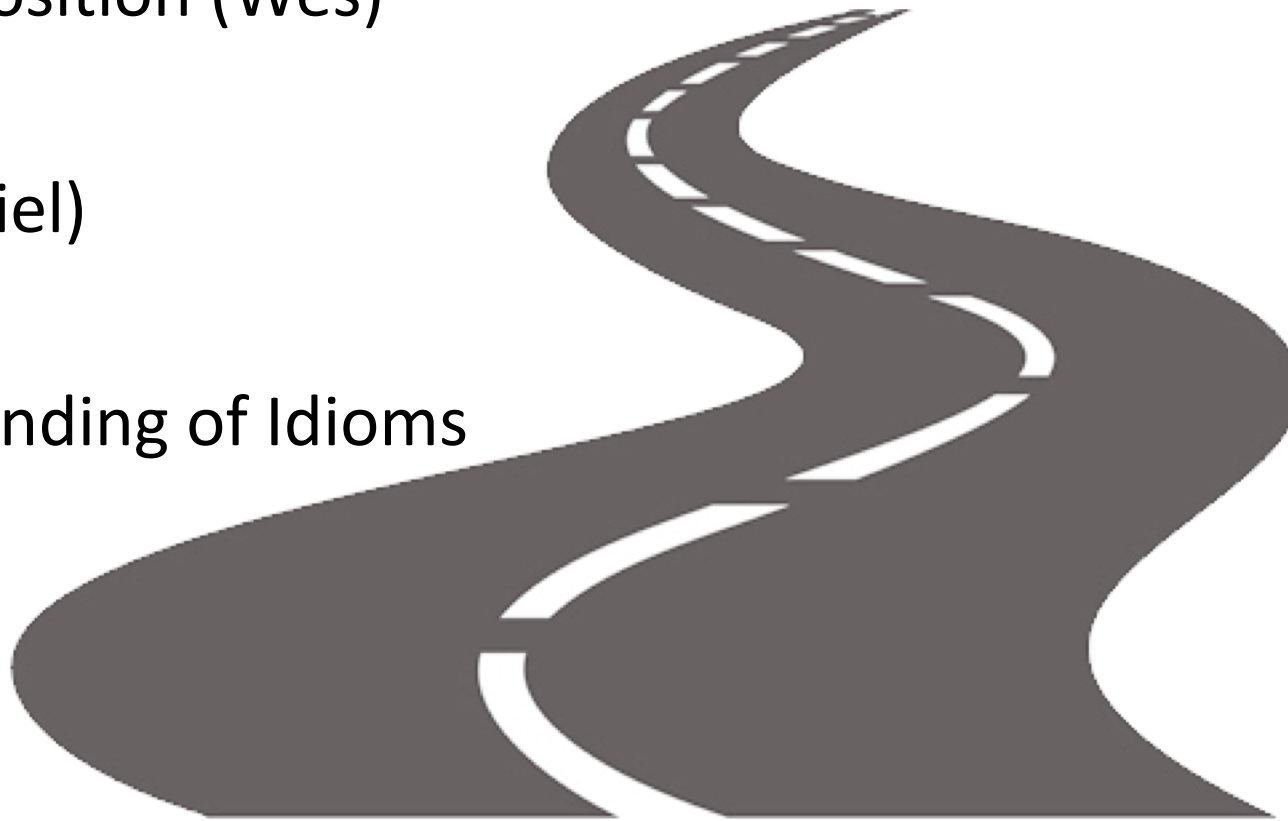
Ling 575 Group 1:

Josh Tanner, Paige Finkelstein, Wes Rose, Elena Khasanova, and Daniel Campos

February 20th, 2020

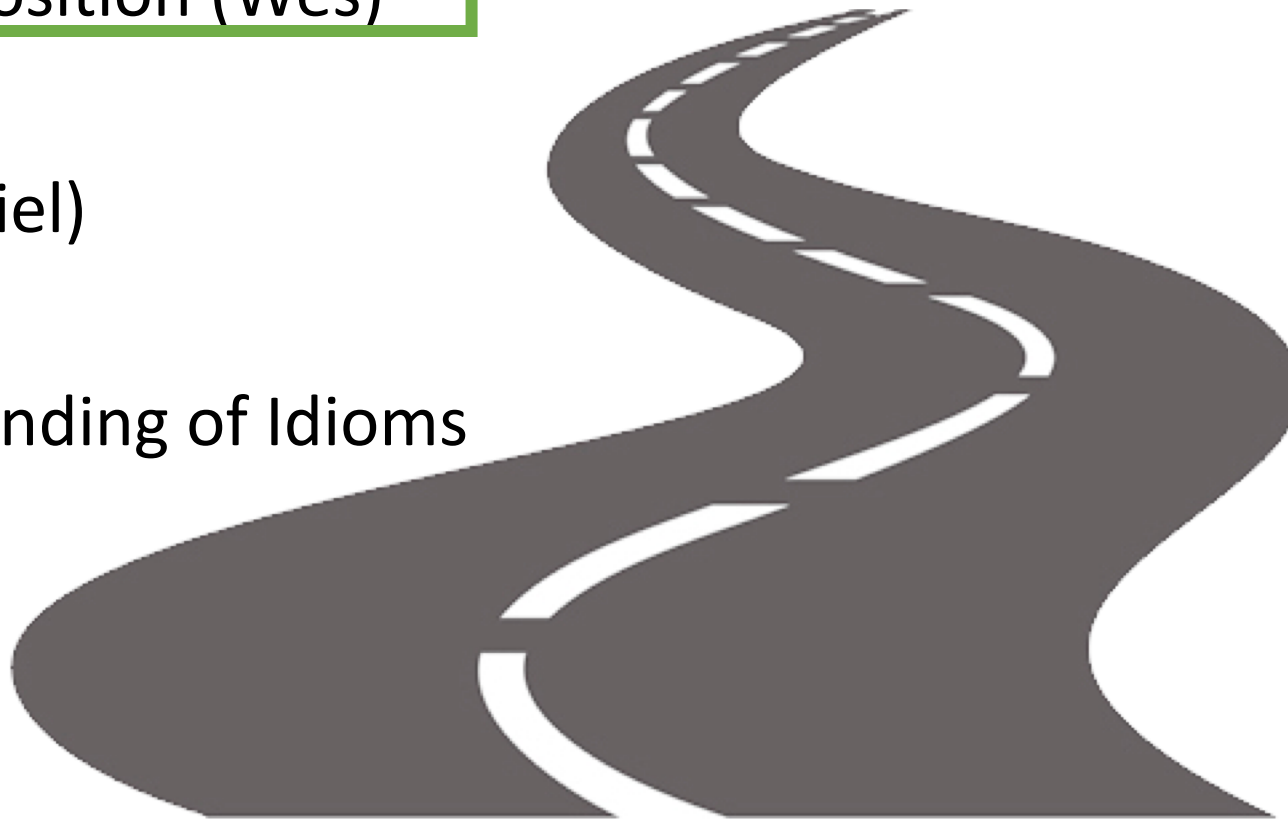
Roadmap

- Overall Introduction
- Evaluating NLM and Lexical Composition (Wes)
- Q&A
- Idioms and Neural Networks (Daniel)
- Q&A
- Our Group Project- NLM Understanding of Idioms
- Q&A



Roadmap

- Overall Introduction
- Evaluating NLM and Lexical Composition (Wes)
- Q&A
- Idioms and Neural Networks (Daniel)
- Q&A
- Our Group Project- NLM Understanding of Idioms
- Q&A



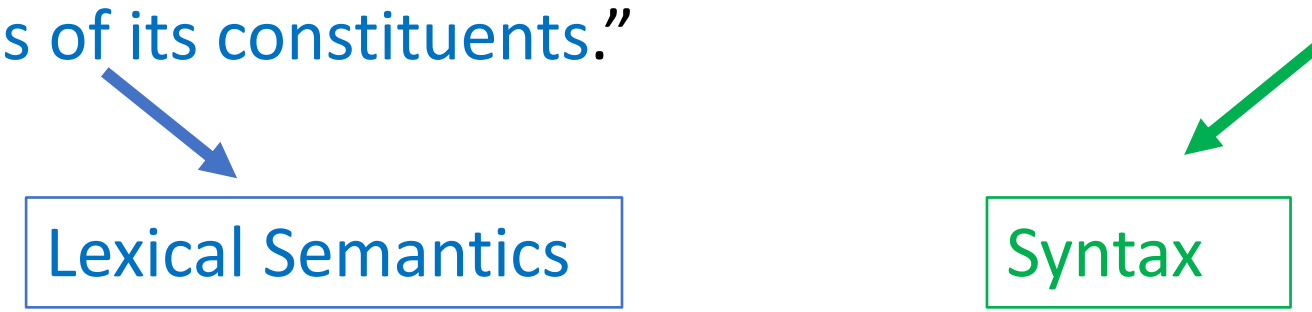
Still a Pain in the Neck: Evaluating Text Representations on Lexical Composition

Vered Shwartz
Computer Science Department, Bar-Ilan University, Ramat-Gan, Israel
vered1986@gmail.com

Ido Dagan
dagan@cs.biu.ac.il

The Principle of Compositionality

- "The meaning of a complex expression is determined by its structure and the meanings of its constituents."



- Given any complex expression *e* in a language *L,* lexical semantics and syntax determine the semantics of *e.*

https://plato.stanford.edu/entries/compositionality/

The Principle of Compositionality

- "The meaning of a complex expression is determined by its structure and the meanings of its constituents."
- Given any complex expression *e* in a language *L,* lexical semantics and syntax determine the semantics of *e.*

Is this always true?

https://plato.stanford.edu/entries/compositionality/

Difficulties with Compositionality

Keep Calm and Carry On ?

- to cause to remain in a given place, situation, or condition

- free from agitation, excitement, or disturbance

- to move while supporting
- to convey by direct communication
- to contain and direct the course of

- used as a function word to indicate the location of something
- used as a function word to indicate a source of attachment or support
- used as a function word to indicate a time frame during which something takes place
- used as a function word to indicate manner of doing something

Example from Schwartz et al.
Definitions from m-w.com

Difficulties with Compositionality

The tea is heating up

To become warm or hot

The argument is heating up

To excite

Which meaning to select?

Example from Schwartz et al.
Definitions from m-w.com

Difficulties with Compositionality

- Meaning Shift
  - The meaning of the phrase departs from the meaning of its constituent words
  - E.g. Carry on, guilt trip, pain in the neck
  - Common in **multi-word expressions**

- Implicit meaning
  - A meaning resulting from composition that requires world knowledge
  - E.g. hot argument vs. hot tea, olive oil vs. baby oil.

Schwartz et al.

Difficulties with Compositionality

- Meaning Shift
  - The meaning of the phrase departs from the meaning of its constituent words
  - E.g. Carry on, guilt trip, pain in the neck

- Implicit meaning
  - A meaning resulting from composition that requires world knowledge
  - E.g. hot argument vs. hot tea, olive oil vs. baby oil.

How do you think Neural Networks will handle these?

Schwartz et al.

Still a Pain in the Neck:
Evaluating Text Representations on Lexical Composition

Vered Shwartz **Ido Dagan**
Computer Science Department, Bar-Ilan University, Ramat-Gan, Israel
vered1986@gmail.com dagan@cs.biu.ac.il

Goals of the paper:

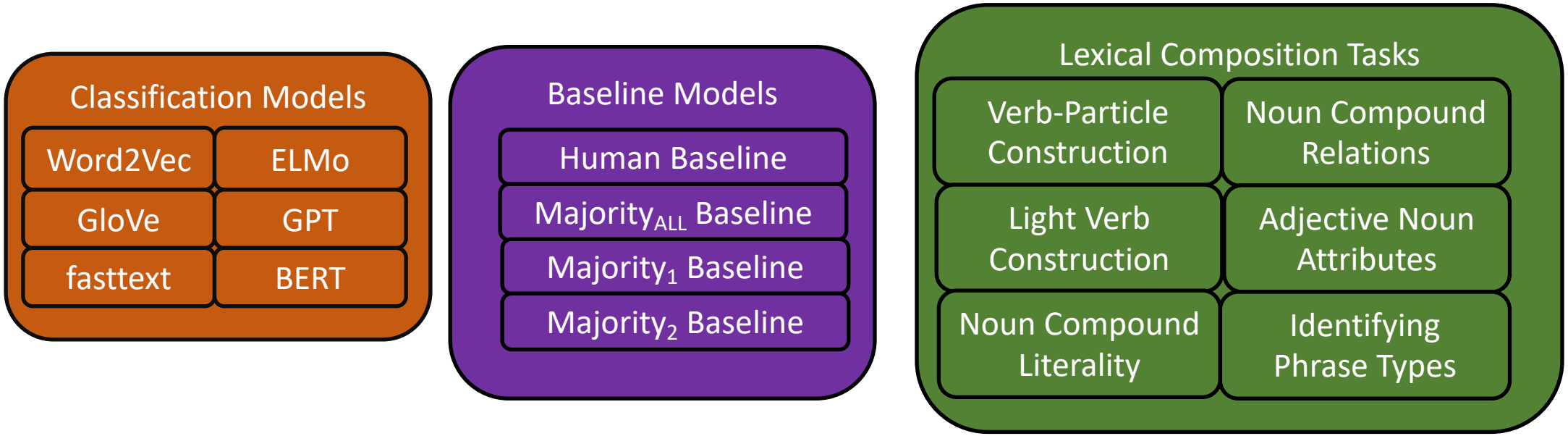
1) Define an evaluation suite for lexical composition for NLP models
   - Based on meaning shift and implicit meaning
2) Evaluate some common word representations using this suite
   - Word2Vec, GloVe, fasttext, ELMo, OpenAI GPT, BERT

Food for Thought

- Would you expect Neural Networks to do better with Meaning Shift or Implicit Meaning?
- What do you think of the tasks that were chosen? Should any tasks be added or expanded?
- How can we improve NLP applications to handle these phenomena?
  - (How do humans handle them?)

Overview of Methodology

- Train 6 classification models, one for each of 6 types of word representations
- For 6 tasks, test each of these models. Compare to each other and to baselines

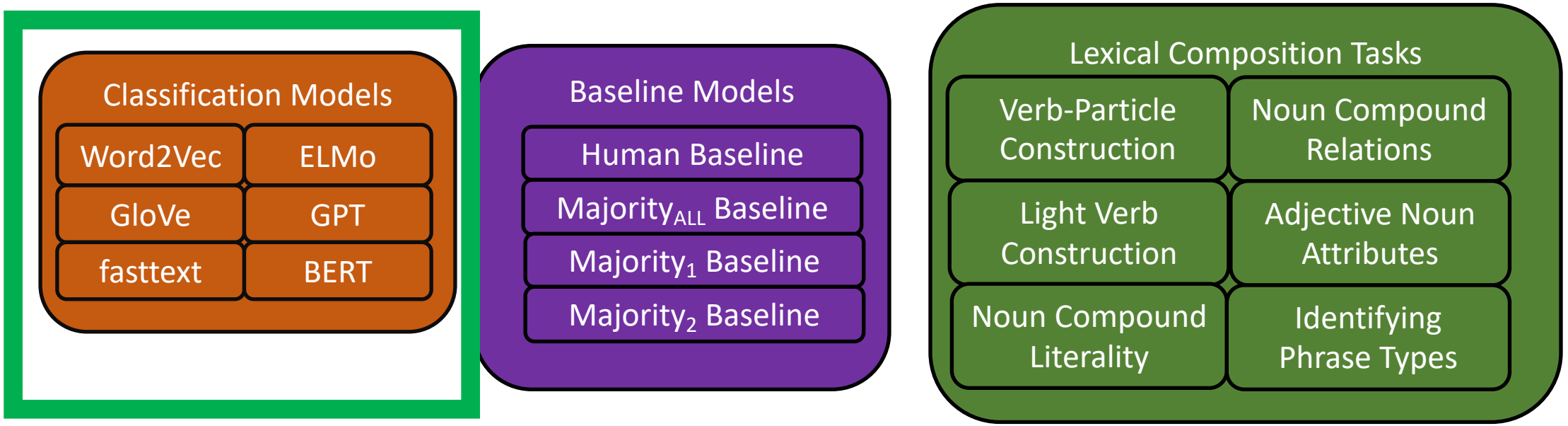


Overview of Methodology

		Task					
		Verb-Particle Construction	Light Verb Construction	Noun Compound Literality	Noun Compound Relations	Adjective Noun Attributes	Identifying Phrase Types
Classification Model	Word2Vec						
	GloVe						
	fasttext						
	ELMo						
	GPT						
	BERT						
	Human Baseline						
	Majority_ALL						
	Majority_1						
	Majority_2						

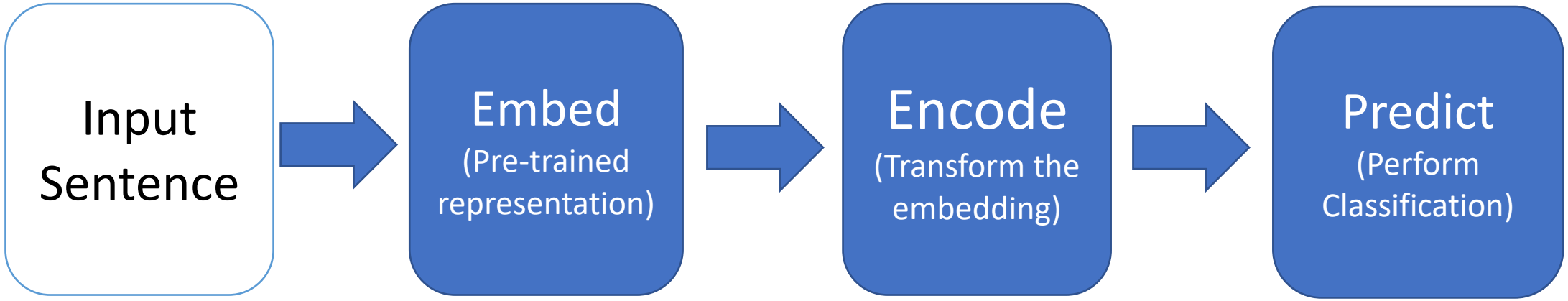
Overview of Methodology

- Train 6 classification models, one for each of 6 types of word representations
- For 6 tasks, test each of these models. Compare to each other and to baselines



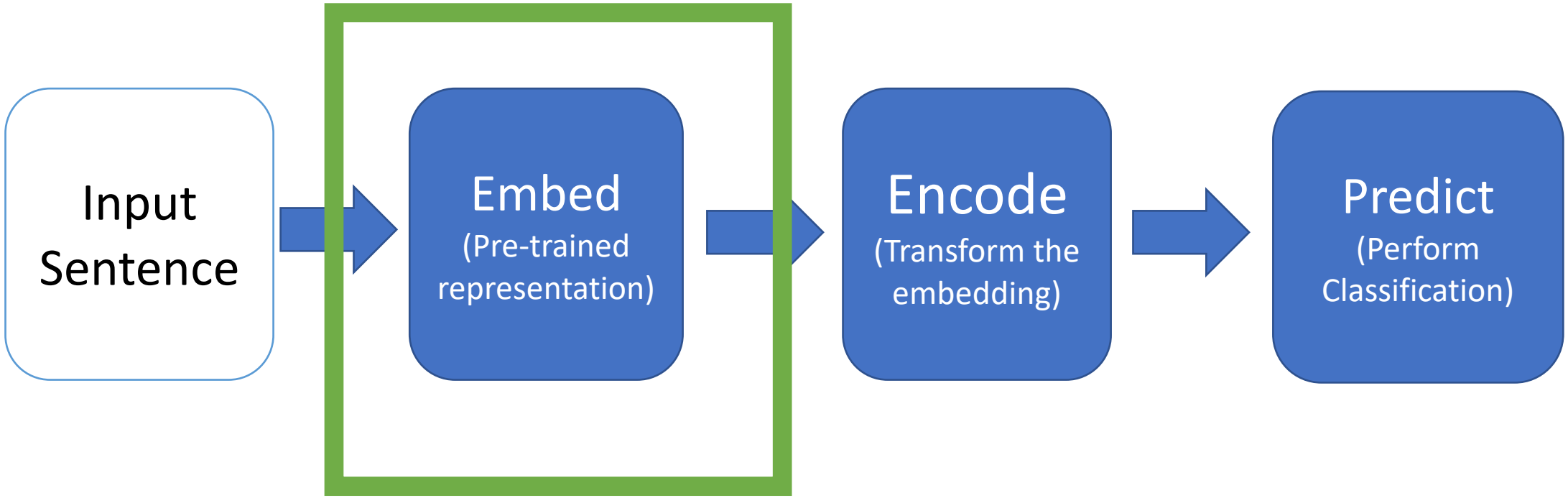
Classification Models

- Embed-Encode-Predict



Classification Models

- Embed-Encode-Predict



Classification Model: Embed (Word Representations)

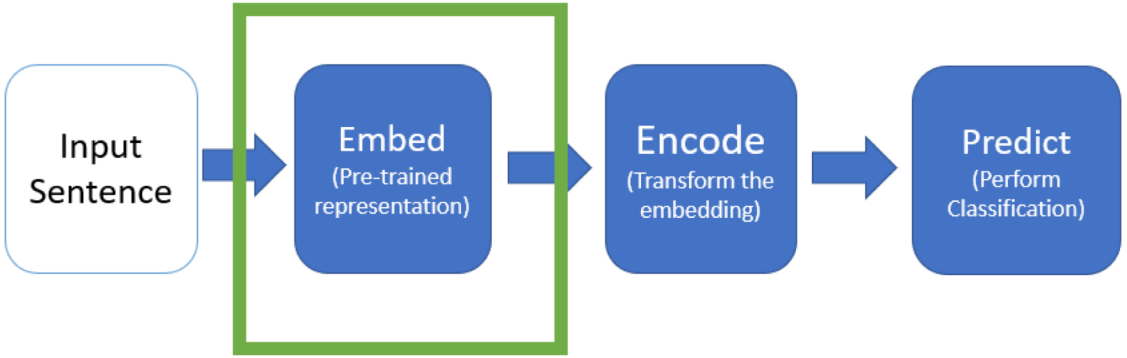
Global Embeddings

- Word2Vec
  - Using Skip-Gram
- GloVe
- fasttext

Contextual Embeddings

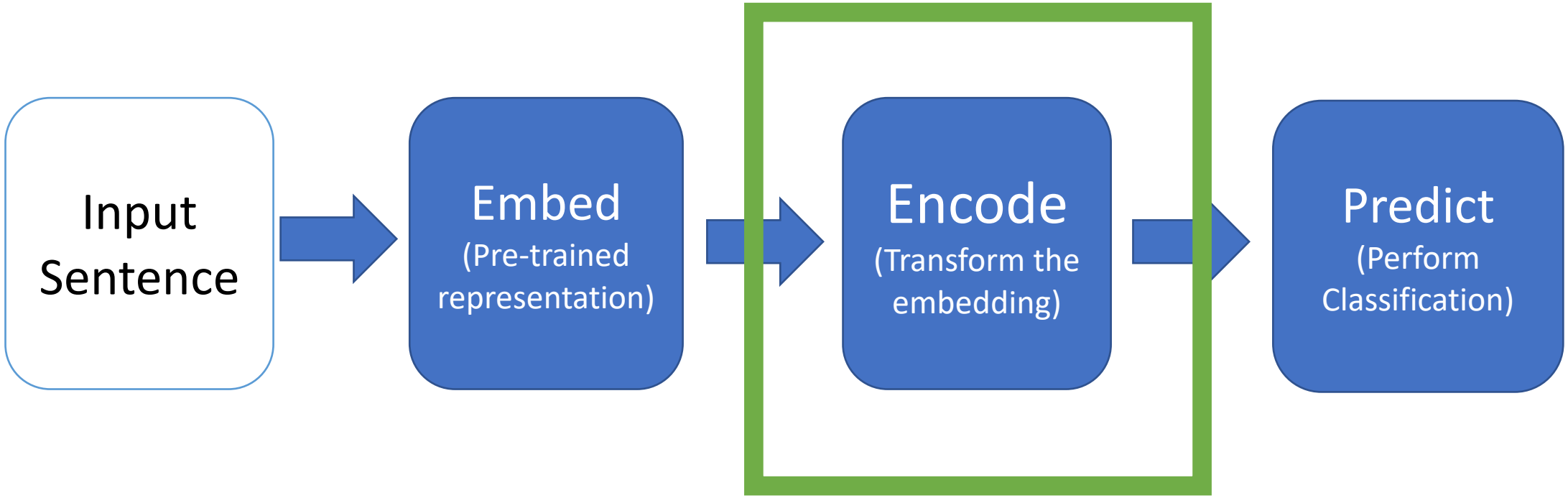
- ELMo
- OpenAI GPT
- BERT

(Use top layer or scalar mix)



Classification Models

- Embed-Encode-Predict



Classification Model: Encode

Input to encode layer is sequence of pretrained embeddings V = <v1,…,vn>

Output is U = <u_1, …, u_n>

biLM

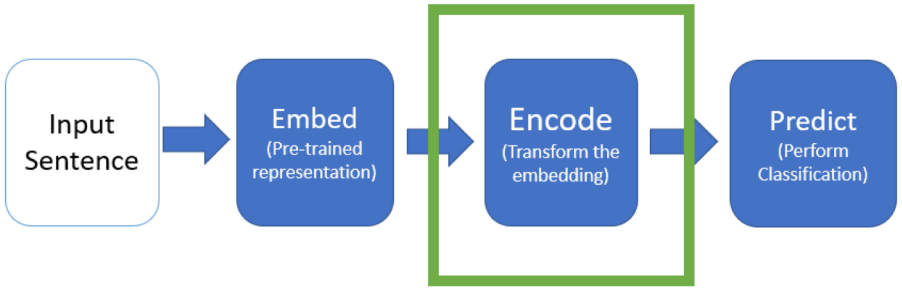
- Encode embedded sequence using biLSTM
- U = biLSTM(V)

Att

- Encode embedded sequence using self-attention
- $U_i = [v_i \Sigma a_{i,j} . V_j]$

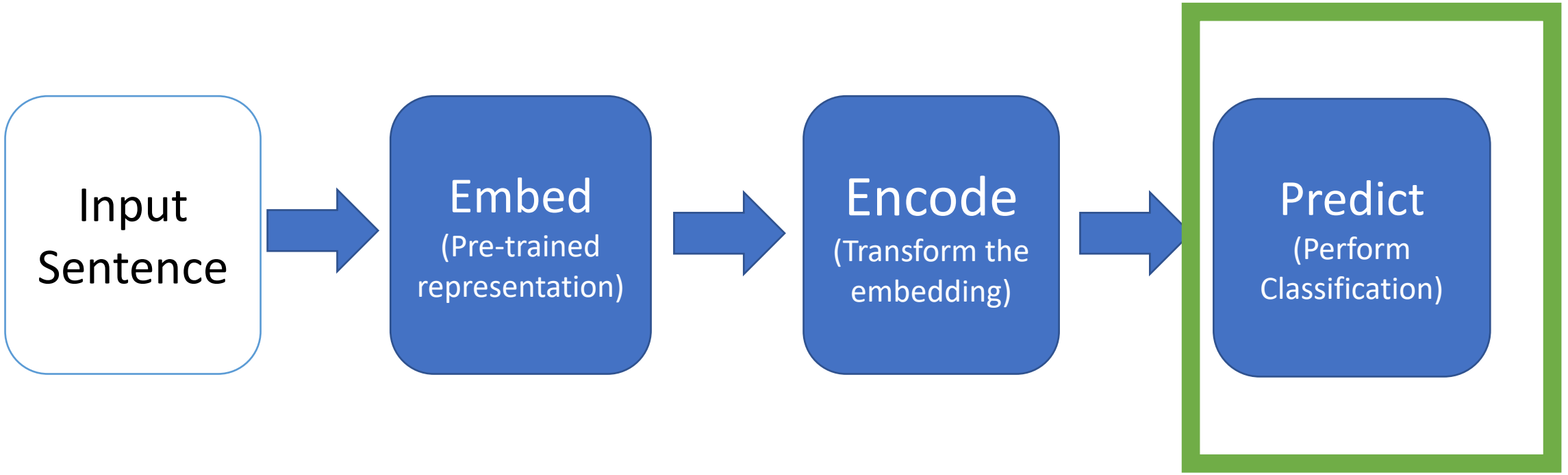
None

- Don't encode the embedded text
- Use the embeddings as they are
- U = U



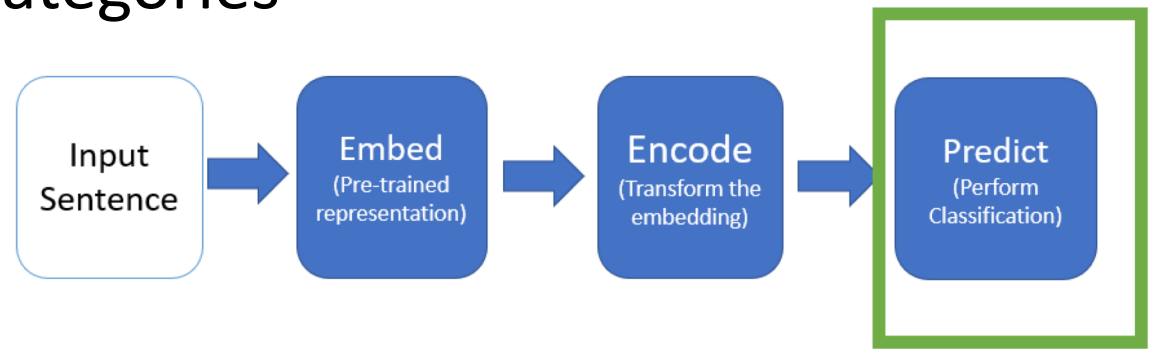
Classification Models

- Embed-Encode-Predict



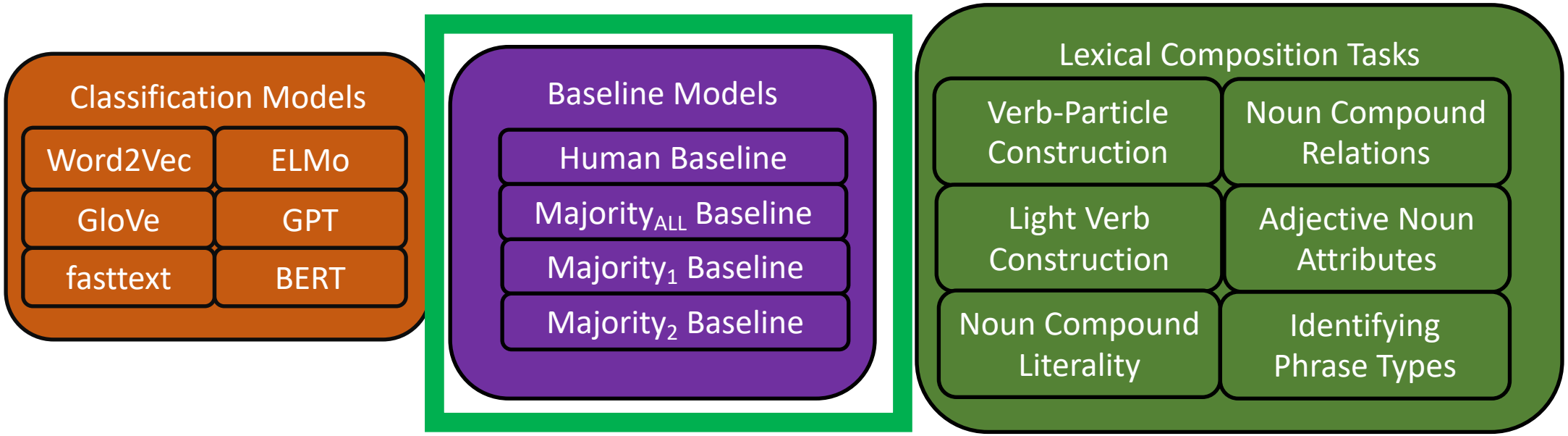
Classification Model: Predict

- Takes output U from Encode layer, and passes it to a feed-forward Neural Network Classifier
- Represent a "span" of text by concatenating end-point vectors
  - E.g. $u_{i,\ldots,i+k} = [u_i ; u_{i+k}]$
- $X = [u_i;u_{i+k};u'_1;u'_l]$
  - u'_1 and u'_l may be empty. For some tasks, a 2nd span is needed.
- X is passed into classifier
- Classifier output is a softmax over all categories



Overview of methodology

- Train 6 classification models, one for each of 6 types of word representations
- For 6 tasks, test each of these models. Compare to each other and to baselines

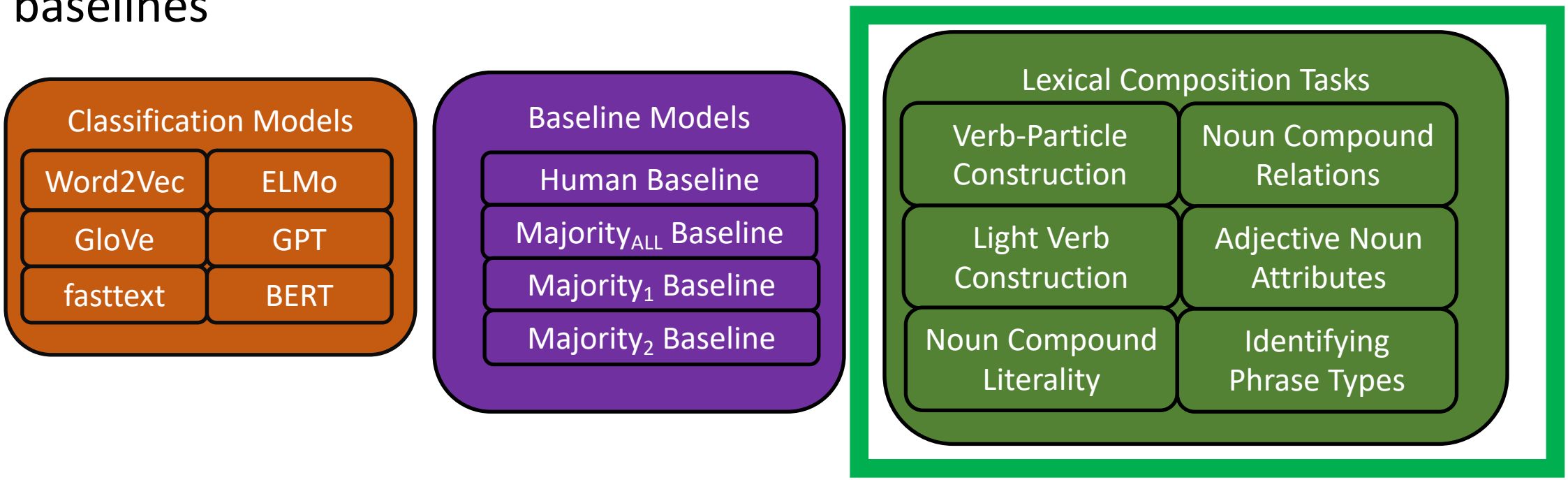


Baselines

Human Baseline	**Majority Baselines**		
• Used Amazon Mechanical Turk • Classified 100 examples for each task • Worker agreement of 80% - 87%	• Majority_{ALL} • Assign most common label in training set to all test items	• Majority_{1} • For each test item, assign most common label in the training set for items with same 1st constituent	• Majority_{2} • For each test item, assign label based on final constituent

Overview of methodology

- Train 6 classification models, one for each of 6 types of word representations
- For 6 tasks, test each of these models. Compare to each other and to baselines



Lexical Composition Tasks

Task Name	Meaning Shift?	Implicit Meaning?
Verb-Particle Construction	X	
Light Verb Construction	X	
Noun Compound Literality	X	
Noun Compound Relations		X
Adjective Noun Attributes		X
Identifying Phrase Type	X	X

Lexical Composition Tasks

Task Name	Meaning Shift?	Implicit Meaning?
Verb-Particle Construction	X	
Light Verb Construction	X	
Noun Compound Literality	X	
Noun Compound Relations		X
Adjective Noun Attributes		X
Identifying Phrase Type	X	X

Task 1: Verb Particle Construction

Given a (verb, preposition) pair from a sentence, is it a verb particle construction?
(Is the verb's meaning changed by the preposition?)

<u>Dataset:</u>
1,348 tagged sentences from the BNC

Example Sentence	Is Verb Particle Construction?
How many Englishmen **gave in** to their emotions like that ?	Yes
It is just this denial of anything beyond what is directly **given in** experience that marks Berkeley out as an empiricist .	No

Tu and Roth 2012

Task 1: Verb Particle Construction

	VPC Classification (Acc)	LVC Classification (Acc)	NC Literality (Acc)	NC Relations (Acc)	AN Attributes (Acc)	Phrase Type (F1)
Majority Baseline	23.6					
Best Global Embedding	60.5					
Best Contextual Embedding	90.0					
Human Baseline	93.8					

Lexical Composition Tasks

Task Name	Meaning Shift?	Implicit Meaning?
Verb-Particle Construction	X	
Light Verb Construction	X	
Noun Compound Literality	X	
Noun Compound Relations		X
Adjective Noun Attributes		X
Identifying Phrase Type	X	X

Task 2: Light Verb Construction

Can the meaning of the verb-noun construction be derived primarily from the meaning of its noun object?

Dataset:
2,162 tagged sentences from the BNC

Example Sentence	Is Light Verb Construction?
I've arranged for you to **have a look** at his file in our library.	Yes
He **had a look** of childish bewilderment on his face.	No

Tu and Roth 2011

Task 2: Light Verb Construction

	VPC Classification (Acc)	LVC Classification (Acc)	NC Literality (Acc)	NC Relations (Acc)	AN Attributes (Acc)	Phrase Type (F1)
Majority Baseline	23.6	43.7				
Best Global Embedding	60.5	74.6				
Best Contextual Embedding	90.0	82.5				
Human Baseline	93.8	83.8				

Lexical Composition Tasks

Task Name	Meaning Shift?	Implicit Meaning?
Verb-Particle Construction	X	
Light Verb Construction	X	
Noun Compound Literality	X	
Noun Compound Relations		X
Adjective Noun Attributes		X
Identifying Phrase Type	X	X

Task 3: Noun Compound Literality

Given a sentence with a {noun1, noun2} compound, is each of the nouns literal or non-literal?

Dataset:
90 annotated examples from ukWaC[6]
3,096 literal examples from Tratz[2] and the PTB-WSJ

Example Sentence	{n1,n2} are literal?
AND tickets for an air boat ride in the Everglades. Wow! Still on **cloud nine.** [6]	{no, no}
Could you also include your **snail mail** address so I can send you a 1999 New Zealand Calendar in Appreciation?[1]	{no, yes}

[6] Reddy et al. 2011
[7] Tratz 2011
[5]ukWaC

Task 3: Noun Compound Literality

	VPC Classification (Acc)	LVC Classification (Acc)	NC Literality (Acc)	NC Relations (Acc)	AN Attributes (Acc)	Phrase Type (F1)
Majority Baseline	23.6	43.7	72.5			
Best Global Embedding	60.5	74.6	80.4			
Best Contextual Embedding	90.0	82.5	91.3			
Human Baseline	93.8	83.8	91.0			

Lexical Composition Tasks

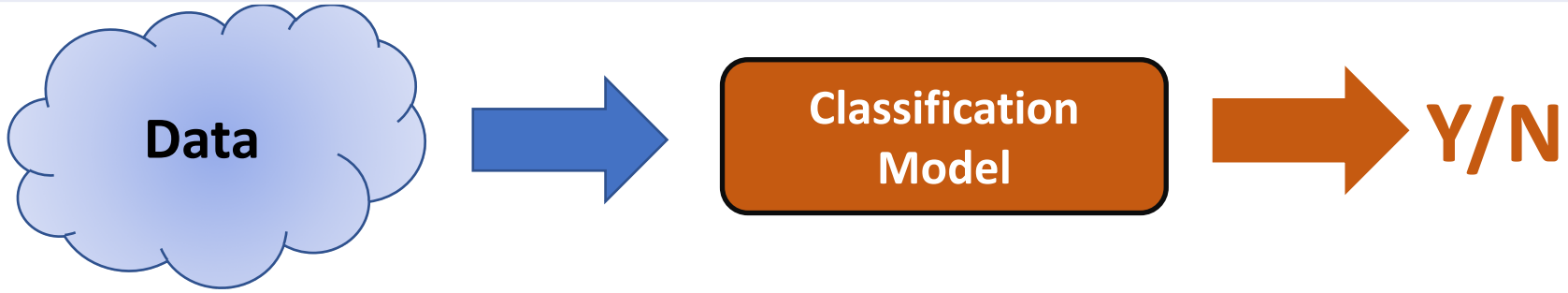
Task Name	Meaning Shift?	Implicit Meaning?
Verb-Particle Construction	X	
Light Verb Construction	X	
Noun Compound Literality	X	
Noun Compound Relations		X
Adjective Noun Attributes		X
Identifying Phrase Type	X	X

Task 4: Noun Compound Relations

Given a sentence with a {noun1, noun2} compound and a paraphrase p, does p describe the semantic relation between noun1 and noun2?

Dataset:
From SemEval 2013[10]: 356 Noun-Compound, annotated with 12,446 paraphrases.

Example Sentence	Valid paraphrase?
{Vietnam has a US$900 million trade surplus in car parts, totaling US$4.4 billion of **car part** exports; **replacement part bought for car**}	Yes
{an appendage (or outgrowth) is an external **body part**, or natural prolongation, that protrudes from an organism's body **; replacement part bought for body**}	no



[10] Hendrickx et al., 2013

Task 4: Noun Compound Relations

	VPC Classification (Acc)	LVC Classification (Acc)	NC Literality (Acc)	NC Relations (Acc)	AN Attributes (Acc)	Phrase Type (F1)
Majority Baseline	23.6	43.7	72.5	50.0		
Best Global Embedding	60.5	74.6	80.4	51.2		
Best Contextual Embedding	90.0	82.5	91.3	54.3		
Human Baseline	93.8	83.8	91.0	77.8		

Lexical Composition Tasks

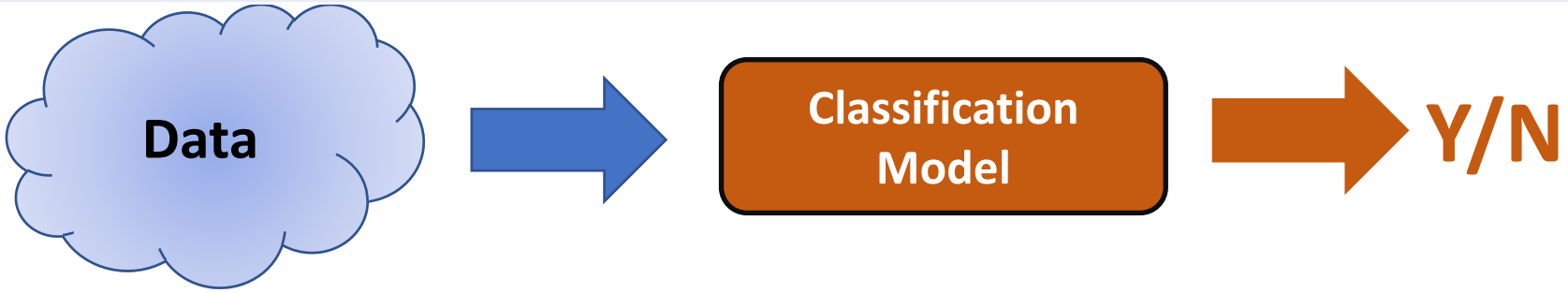
Task Name	Meaning Shift?	Implicit Meaning?
Verb-Particle Construction	X	
Light Verb Construction	X	
Noun Compound Literality	X	
Noun Compound Relations		X
Adjective Noun Attributes		X
Identifying Phrase Type	X	X

Task 5: Adjective Noun Attributes

Given a sentence s with Adjective-Noun combination AN paired with an attribute AT: Is AT implicitly conveyed in AN?

Dataset:
HeiPLAS[8] with 1,589 annotated examples from WordNet

Example Sentence	Is AT attribute of AN?
{Heat traps are valves or loops of pipe installed on the cold water inlet and **hot water** outlet pipes on water heaters, **temperature**}	Yes
{A **hot argument** takes place between Sanjana and her father, and she runs away to Charan, **temperature**}	No



[8]Hartung 2015

Task 5: Adjective Noun Attributes

	VPC Classification (Acc)	LVC Classification (Acc)	NC Literality (Acc)	NC Relations (Acc)	AN Attributes (Acc)	Phrase Type (F1)
Majority Baseline	23.6	43.7	72.5	50.0	50.0	
Best Global Embedding	60.5	74.6	80.4	51.2	53.8	
Best Contextual Embedding	90.0	82.5	91.3	54.3	65.1	
Human Baseline	93.8	83.8	91.0	77.8	86.4	

Lexical Composition Tasks

Task Name	Meaning Shift?	Implicit Meaning?
Verb-Particle Construction	X	
Light Verb Construction	X	
Noun Compound Literality	X	
Noun Compound Relations		X
Adjective Noun Attributes		X
Identifying Phrase Type	X	X

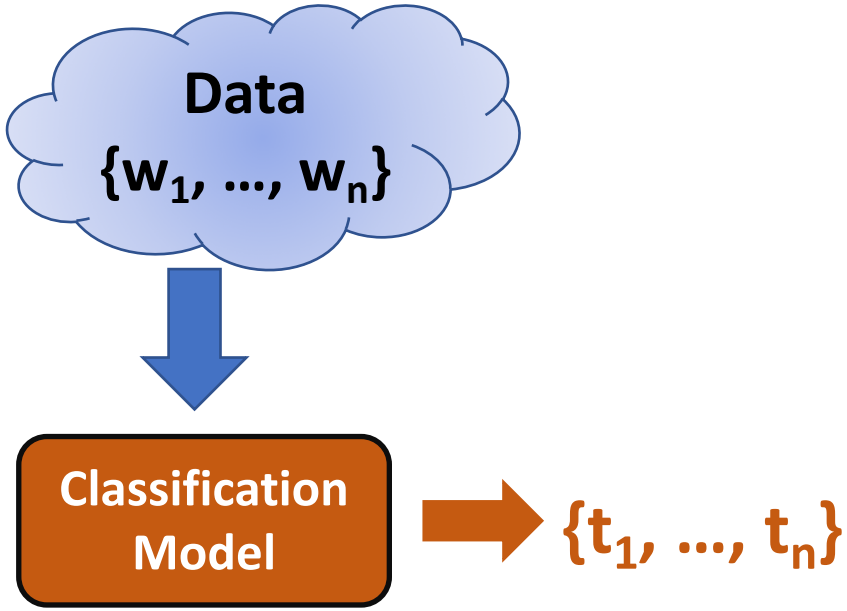
Task 6: Identifying Phrase Type

Given a sentence s with words $\{w_1, w_2, \ldots, w_n\}$, output a sequence of BIO labels for each word w_i. For each word w_i: is it part of a phrase, and if so what is the phrase type?

Dataset:
STREUSEL corpus[9] based on reviews section of English Web Treebank

Types of Phrases

[2]Sorted by frequency: noun phrase, weak (compositional) MWE, verb-particle construction, verbal idioms, prepositional phrase, auxiliary, adposition, discourse / pragmatic expression, inherently adpositional verb, adjective, determiner, adverb, light verb construction, non-possessive pronoun, full verb or copula, conjunction.



[9]Schneider and Smith 2015

Task 6: Identifying Phrase Type

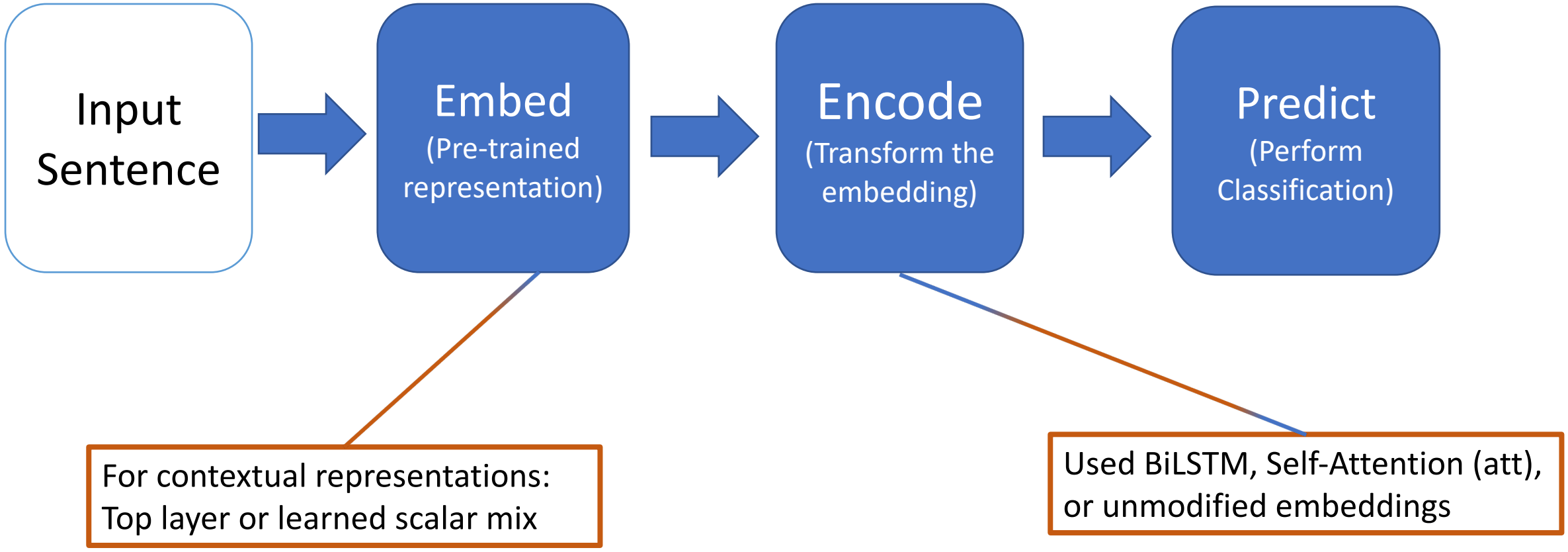
	VPC Classification (Acc)	LVC Classification (Acc)	NC Literality (Acc)	NC Relations (Acc)	AN Attributes (Acc)	Phrase Type (F1)
Majority Baseline	23.6	43.7	72.5	50.0	50.0	26.6
Best Global Embedding	60.5	74.6	80.4	51.2	53.8	44.0
Best Contextual Embedding	90.0	82.5	91.3	54.3	65.1	64.8
Human Baseline	93.8	83.8	91.0	77.8	86.4	

Model Performance on Two Phenomena

	Meaning Shift			Implicit Meaning		Both
	VPC Classification (Acc)	**LVC Classification (Acc)**	**NC Literality (Acc)**	**NC Relations (Acc)**	**AN Attributes (Acc)**	**Phrase Type (F1)**
Majority Baseline	23.6	43.7	72.5	50.0	50.0	26.6
Best Global Embedding	60.5	74.6	80.4	51.2	53.8	44.0
Best Contextual Embedding	90.0	82.5	91.3	54.3	65.1	64.8
Human Baseline	93.8	83.8	91.0	77.8	86.4	
Best Model – Human Baseline	**-3.8**	**-1.3**	**.3**	**-23.5**	**-21.3**	

Extra Analysis Tasks
(If Time)

Best Encodings and Layers



Best Encodings and Layers

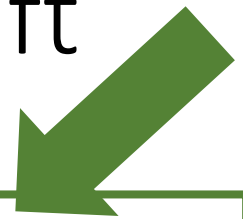
Model	VPC Classification		LVC Classification		NC Literality		NC Relations		AN Attributes		Phrase Type	
	Layer	Encoding	Layer	Encoding	Layer	Encoding	Layer	Encoding	Layer	Encoding	Layer	Encoding
ELMo	All	Att	All	biLM	All	Att/None	**Top**	**biLM**	All	None	All	biLM
OpenAI GPT	All	None	**Top**	**Att/None**	Top	None	All	biLM	Top	None	All	biLM
BERT	**All**	**Att**	All	biLM	**All**	**Att**	All	None	**All**	**None**	**All**	**biLM**

Table 5: The best setting (layer and encoding) for each contextualized word embedding model on the various tasks. Bold entries are the best performers on each task.

Model	VPC Classification	LVC Classification	NC Literality	NC Relations	AN Attributes	Phrase Type
word2vec	biLM	Att	biLM/Att	None	-	None/biLM
GloVe	**biLM**	**Att**	**Att**	biLM	-	**biLM**
fastText	Att	biLM	biLM	**biLM**	**Att**	biLM

Table 6: The best encoding for each word embedding model on the various tasks. Bold entries are the best performers on each task. Dash marks no preference.

Analysis of Meaning Shift



	Meaning Shift			Implicit Meaning		Both
	VPC Classification (Acc)	**LVC Classification (Acc)**	**NC Literality (Acc)**	**NC Relations (Acc)**	**AN Attributes (Acc)**	**Phrase Type (F1)**
Majority Baseline	23.6	43.7	72.5	50.0	50.0	26.6
Best Global Embedding	60.5	74.6	80.4	51.2	53.8	44.0
Best Contextual Embedding	90.0	82.5	91.3	54.3	65.1	64.8
Human Baseline	93.8	83.8	91.0	77.8	86.4	
Best Model – Human Baseline	**-3.8**	**-1.3**	**.3**	**-23.5**	**-21.3**	

Meaning Shift: Verb-Particle Classification

	VPC Classification (Acc)
Majority Baseline	23.6
Best Global Embedding	60.5
Best Contextual Embedding	90.0
Human Baseline	93.8
Best Model – Human Baseline	**-3.8**

Best Performer: BERT + All + Att

Do BERT embeddings really have all of the information necessary?

Ablation Task:

- Choose several ambiguous verb-preposition pairs
- Compute BERT representation for each example of each pair
- Project representations into 2D Space using t-SNE

Meaning Shift: Verb-Particle Classification

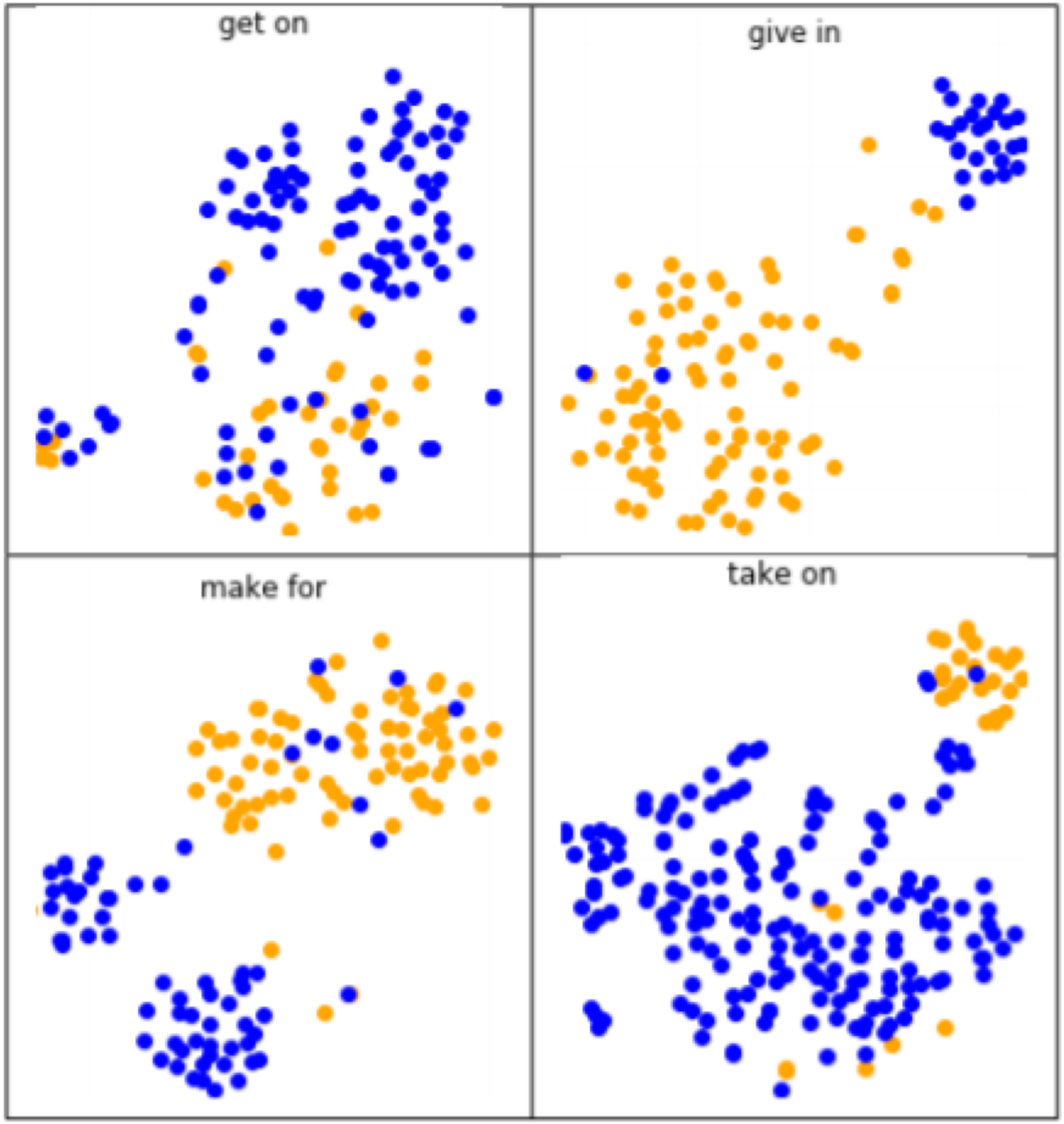
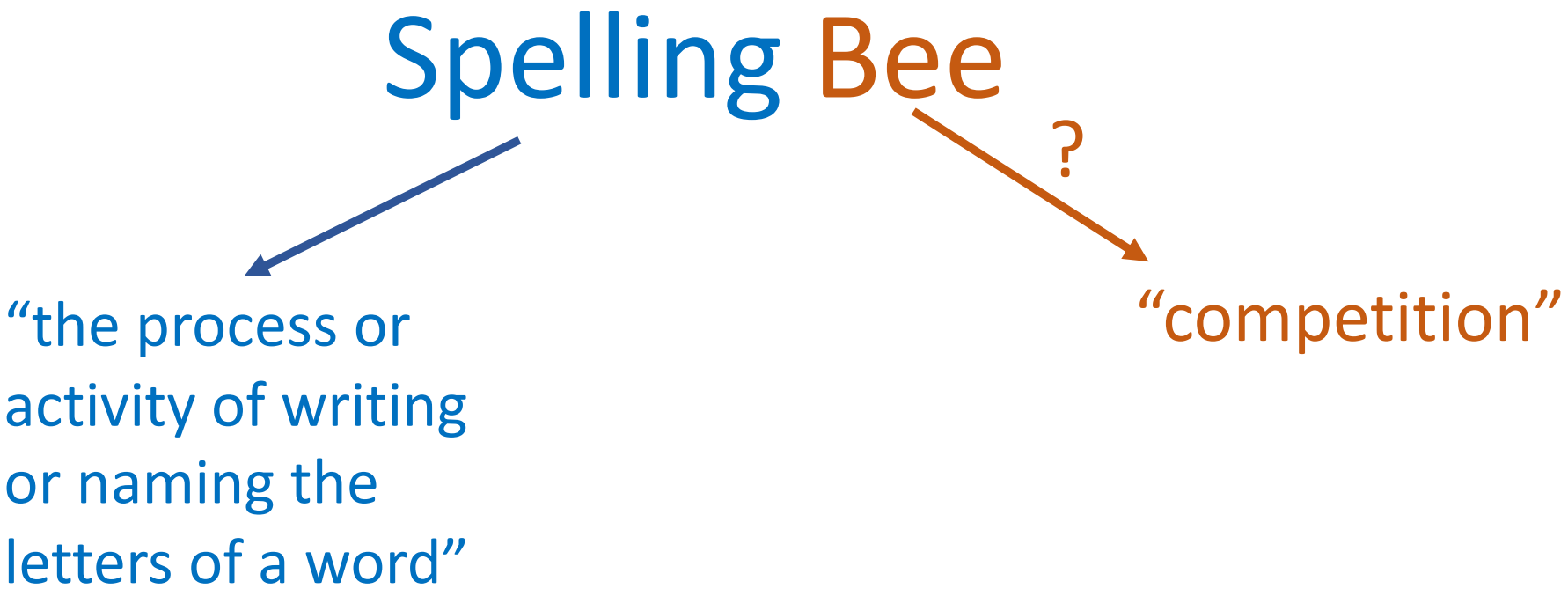


Figure 2: t-SNE projection of BERT representations of verb-preposition candidates for VPC. Blue (dark) points are positive examples and orange (light) points are negative.

Meaning Shift: Non-literality as Rare Sense



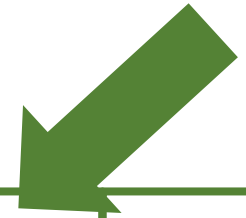
Meaning Shift: Non-literality as Rare Sense

- Can word embeddings be used for “word sense induction?”
- Sample target words that appear in literal and non-literal examples
- Use contextualized word embeddings in these examples to predict best substitute for target word

Meaning Shift: Non-literality as Rare Sense

ELMo		OpenAI GPT		BERT		ELMo		OpenAI GPT		BERT	
The Queen and her husband were on a train [trip]$_L$ from Sydney to Orange.						Creating a guilt [trip]$_N$ in another person may be considered to be psychological manipulation in the form of punishment for a perceived transgression.					
ride	1.24%	to	0.02%	**travelling**	19.51%	tolerance	0.44%	that	0.03%	**reaction**	8.32%
carriage	1.02%	**headed**	0.01%	**running**	8.20%	**fest**	0.23%	so	0.02%	**feeling**	8.17%
journey	0.73%	**heading**	0.01%	**journey**	7.57%	avoidance	0.16%	**trip**	0.01%	attachment	8.12%
heading	0.72%	that	0.009%	**going**	6.56%	onus	0.15%	he	0.01%	**sensation**	4.73%
carrying	0.39%	and	0.005%	**headed**	5.75%	association	0.14%	she	0.01%	note	3.31%
Richard Cromwell so impressed the king with his valour, that he was given a [diamond]$_L$ ring from the king's own finger.						She became the first British monarch to celebrate a [diamond]$_N$ wedding anniversary in November 2007.					
diamond	0.23%	and	0.01%	**silver**	15.99%	customary	0.20%	new	0.11%	royal	1.58%
wedding	0.19%	of	0.01%	**gold**	14.93%	royal	0.17%	british	0.02%	1912	1.23%
pearl	0.18%	to	0.01%	**diamond**	13.18%	sacrifice	0.15%	victory	0.01%	recent	1.10%
knighthood	0.16%	ring	0.01%	**golden**	12.79%	400th	0.13%	french	0.01%	1937	1.08%
hollow	0.15%	in	0.01%	new	4.61%	**10th**	0.13%	royal	0.01%	1902	1.08%

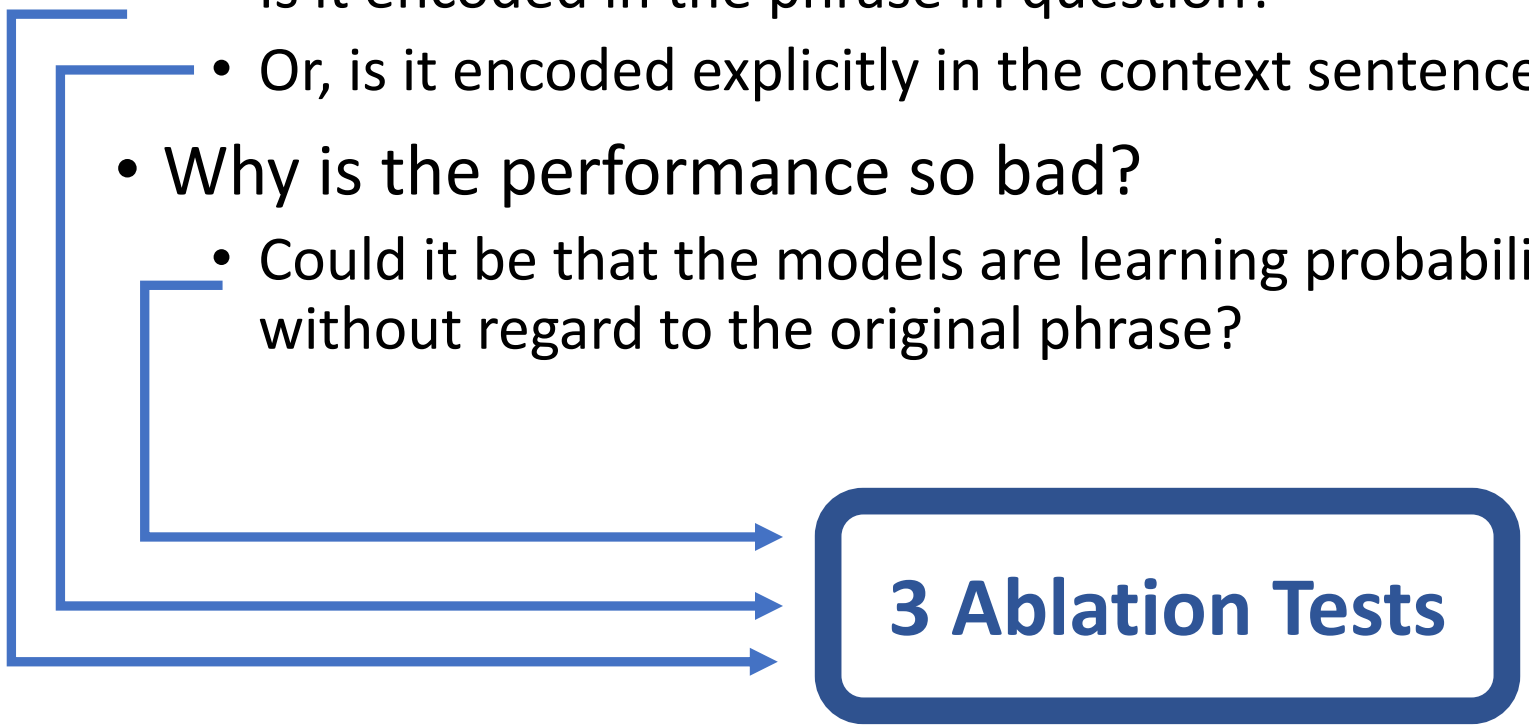
Analysis of Implicit Meaning



	Meaning Shift			Implicit Meaning		Both
	VPC Classification (Acc)	**LVC Classification (Acc)**	**NC Literality (Acc)**	**NC Relations (Acc)**	**AN Attributes (Acc)**	**Phrase Type (F1)**
Majority Baseline	23.6	43.7	72.5	50.0	50.0	26.6
Best Global Embedding	60.5	74.6	80.4	51.2	53.8	44.0
Best Contextual Embedding	90.0	82.5	91.3	54.3	65.1	64.8
Human Baseline	93.8	83.8	91.0	77.8	86.4	
Best Model – Human Baseline	**-3.8**	**-1.3**	**.3**	**-23.5**	**-21.3**	

Analysis of Implicit Meaning

- Where does the knowledge of the implicit meaning originate?
  - Is it encoded in the phrase in question?
  - Or, is it encoded explicitly in the context sentence around the phrase?
- Why is the performance so bad?
  - Could it be that the models are learning probability of paraphrases alone, without regard to the original phrase?



Analysis of Implicit Meaning

3 Ablation Tests

Original Phrase: "Today, the house has become a **wine bar** or bistro called Barokk"

Test 1 (-phrase):

- Mask the phrase in the context sentence
- "Today, the house has become a **something** or bistro called Barokk"

Test 2 (-Context):

- Replace the context sentence with the phrase itself
- **"wine bar"**

Test 3 (-context + phrase):

- Omit the context sentence all together. Provide only the paraphrase
- "bar where people drink wine"

Take each modified context sentence, and evaluate on NC Relations and AN attributes tasks

Analysis of Implicit Meaning

	NC Relations	AN Attributes
Majority	50.0	50.0
-Phrase	50.0	55.66
-Context	45.06	63.21
-(Context+Phrase)	45.06	59.43
Full Model	54.3	65.1

Summary – “Still a pain in the neck”

- Understanding the meanings of phrases is not straightforward
- Meaning Shift and Implicit Meaning
- 6 tasks were developed to evaluate model understanding of these phenomena
- 6 pre-trained language models were evaluated on these tasks
- The models do pretty well with meaning shift. They struggle with implicit meaning

References for pain in the neck

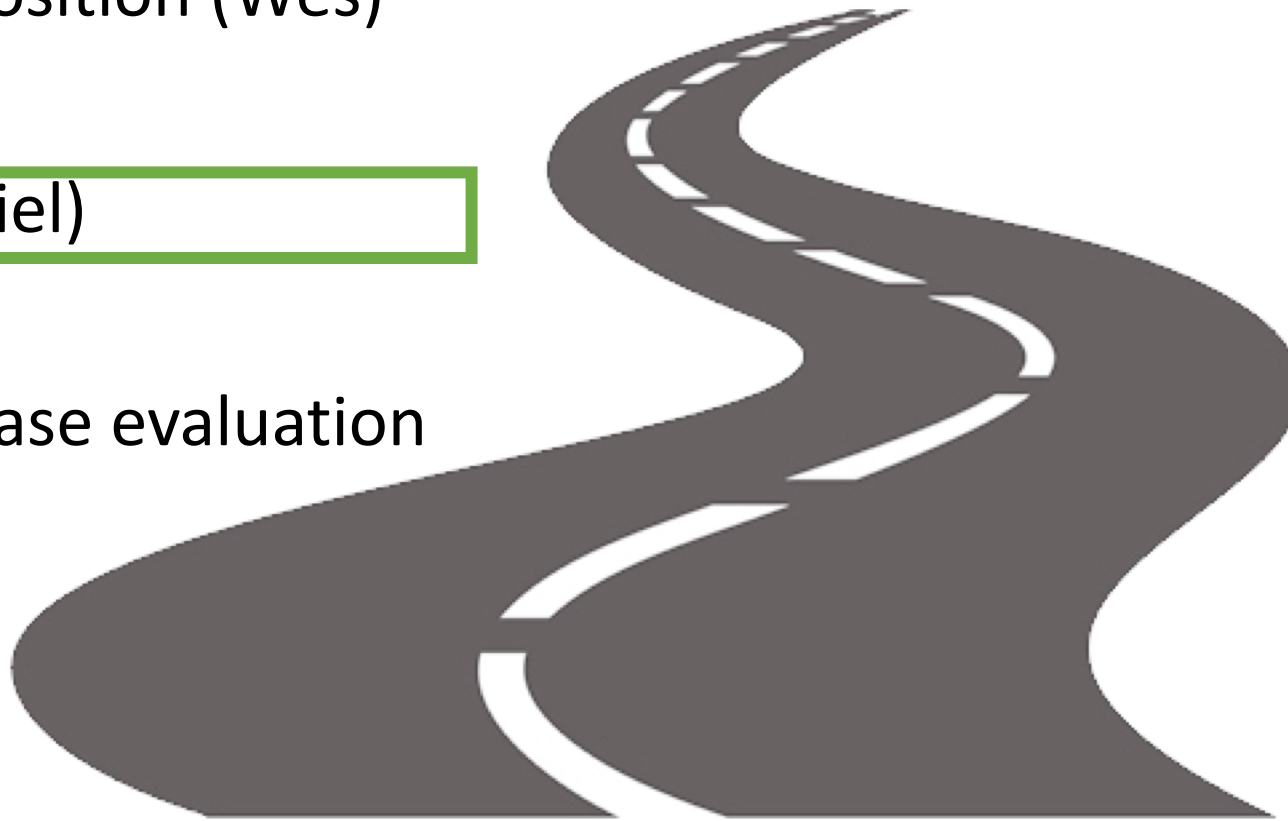
1. Stanford Encyclopedia of Philosophy - https://plato.stanford.edu/entries/compositionality/
2. Merriam-Webster - https://www.merriam-webster.com/dictionary
3. Tu and Roth (2012) - https://www.aclweb.org/anthology/S12-1010/
4. Tu and Roth (2011) - https://www.aclweb.org/anthology/W11-0807/
5. ukWaC corpus - https://www.sketchengine.eu/ukwac-british-english-corpus/
6. Reddy et al. (2011) - https://www.aclweb.org/anthology/I11-1024/
7. Tratz (2011) - http://digitallibrary.usc.edu/cdm/ref/collection/p15799coll3/id/176191
8. Hartung (2015) - https://archiv.ub.uni-heidelberg.de/volltextserver/20013/
9. Schneider and Smith (2015) - https://www.aclweb.org/anthology/N15-1177/
10. Hendrickx et al. (2010) - https://www.aclweb.org/anthology/S13-2025/

Questions / Comments?

- Would you expect Neural Networks to do better with Meaning Shift or Implicit Meaning?
  - **According to the results – meaning shift. Is this surprising?**
- What do you think of the tasks that were chosen? Should any tasks be added or expanded?
  - **Our group is interested in examining idioms more closely**
- How can we improve NLP applications to handle these phenomena?
  - (How do humans handle them?)
  - **For implicit meaning / world knowledge, see Vered Schwartz' Treehouse talk**

Roadmap

- Overall Introduction
- Evaluating NLM and Lexical Composition (Wes)
- Q&A
- Idioms and Neural Networks (Daniel)
- Q&A
- Our Group Project- Idiom paraphrase evaluation
- Q&A



Overview

- Probing task focused on model performance with selective dataset pruning.
- What kind of features in an idiom’s vector are exploited by a NN in classification of idioms vs literals?
- Hypothesis #1: the network could be using the idea of concreteness vs. abstractness to identify idioms as compared to literal phrases.
- Hypothesis #2:the network uses ambiguity as a factor, with the idea being that idioms are more ambiguous on average than literal language.

Definitions

- “metaphors(e.g., my job is a jail) reflect a transparent mapping from concrete examples in a source domain (e.g., the physical confinement of a jail) to the abstract concept in the target domain(e.g., the psychological constraints and tediousness of a job)” - Senaldi et al. 2019
- “idioms (e.g. buy the farm ‘to pass away’, shoot the breeze ‘to chat idly’) synchronically appear as a heterogeneous class of semantically non-compositional multiword units that all exhibit greater lexicosyntactic rigidity, proverbiality and emotional valence with respect to literal expressions.

ELI5 aka Explain like I’m 5

- Metaphors map something real to a non-specific representation in a specific domain.
- Idioms are a broad range of multiword units which map to some literal expression that only work if the complete structure is preserved(e.g. my heart aches != my myocardium hurts)

Related Work

- Neural networks using Word2Vec do well on classification of metaphors. Bizzoni et al. (2017a)
  - In exploring the cosine distance between the nouns and learned metaphorical representation authors found the NN leveraged the concrete-> abstract shift mapping to established linguistic knowledge of metaphors.
- They also do well on classifying idioms! Bizzoni et al. (2017b)
  - In Italian.
  - Entire phrase is treated as one token(e.g. spill the beans is one token)
  - No exploration in what kind of shift is happening in NN
- Idioms, like metaphors, tend to be used to convey abstract concepts and are, generally speaking, less concrete in meaning with respect to literals (Citron et al., 2016)

Guiding Question

- What kind of features in an idiom’s vector are exploited by a NN in classification of idioms vs literals?

Dataset

- 174 Italian and 120 English idiomatic and literal verb noun constructions
  - Italian
    - 87 randomly chosen Italian verbal idioms from idiomatic dictionaries.
      - Extract usage from itWaC corpus
      - Some rare(63 occurrences) other common(15,784)
    - 87 only literal verb phrases selected randomly that matched the idiomatic distribution
  - English
    - 120 VN idiomatic and literal expressions from COCA corpus
    - 60 Idioms and 60 literals following similar procedure to Italian.
- Annotation of correctness and ambiguity
  - Linguistics students/researchers
  - Rating for abstractness/concreteness (1-7)
  - Rating for plausibility of regular usage of VN literal usage (1-7)
  - Rating for ambiguity (1-7)
  - Literals rated as more concrete
    - 4.84 average for Italian literals
    - 3.16 average for Italian idioms
    - 6.20 average for English literals
    - 2.43 average for English idioms

Method

- Train Word2Vec and fastText embeddings on custom corpus and use vectors in a classifier to classify idiom vs literal
- Test model performance with training on various subsets of dataset
  - Trained on complete dataset(with random subsamples)
  - Trained with concrete literals removed from training set.
    - Removed all literals with concreteness > 5
    - If NN relied on differences in concreteness between literals and idioms model performance should greatly drop here
  - Trained with semantically ambiguous idioms removed from training data
    - Removed all idioms with average ambiguity of >5
    - Since idioms can have literal and idiomatic representation model should learn a more varied distribution.

Training

- FastText and Word2Vec trained on itWAC(Italian) and COCA(English)
  - 300 dimensions
  - SkipGram
  - 5 word window
  - 10 negative samples
- Classifier on idiom or not
  - 3 hidden layers
  - 300 -> 12 -> 8 -> 1
  - Sigmoid activation

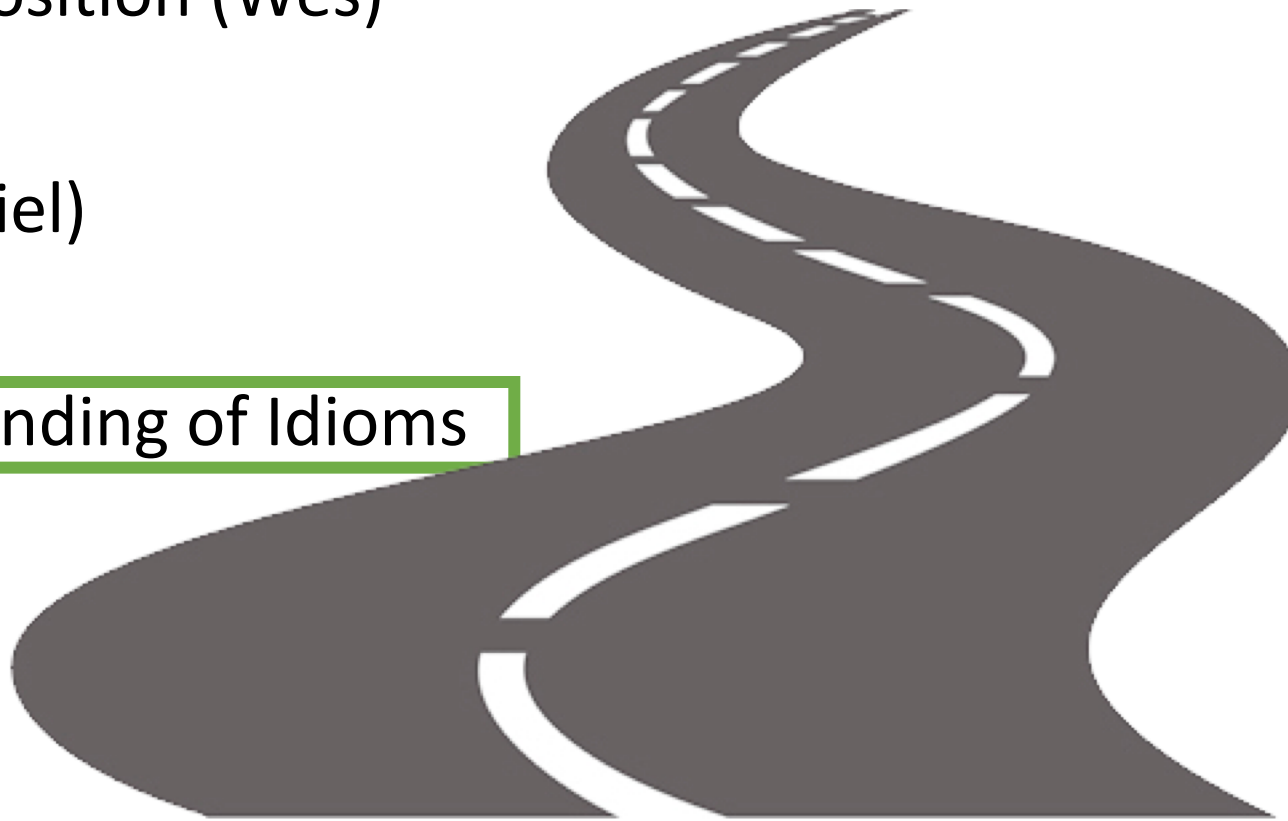
Results

- NN is likely exploiting a difference in concreteness/ambiguity of expression.
- When NN are trained to spot idioms they exploit underlying sematic features.
- Suggest further annotation of idioms to explore what NNs are learning.

Dataset	Model	Avg. training size	Test size	Avg. F1 (Word2vec)	SD	Avg. F1 (fastText)	SD
ITA	TOTAL	140 (70+70)	14 (37+37)	.78	.04	.71	.05
ITA	AMBIGUITY	87.2 (43.6+43.6)	14 (37+37)	.71	.06	.68	.06
ITA	CONCRETENESS	62 (31+31)	14 (37+37)	.41	.1	.40	.13
ITA	RANDOM		14 (37+37)	.44 (SD = .10)			
ENG	TOTAL	96 (48+48)	24 (12+12)	.65	.05	.64	.04
ENG	AMBIGUITY	47.6 (23.8+23.8)	24 (12+12)	.58	.15	.60	.09
ENG	CONCRETENESS	31.2 (15.6+15.6)	24 (12+12)	.33	0	.49	.21
ENG	RANDOM		24 (12+12)	.49 (SD = .10)			

Roadmap

- Overall Introduction
- Evaluating NLM and Lexical Composition (Wes)
- Q&A
- Idioms and Neural Networks (Daniel)
- Q&A
- Our Group Project- NLM Understanding of Idioms
- Q&A



Our Research: NLM Understanding of Idioms

Overview

- Broad research question: can NLMs understand idioms and their underlying meaning?
- What do contextual representations of idioms capture in vector space? Do their approximate the non idiomatic meaning?

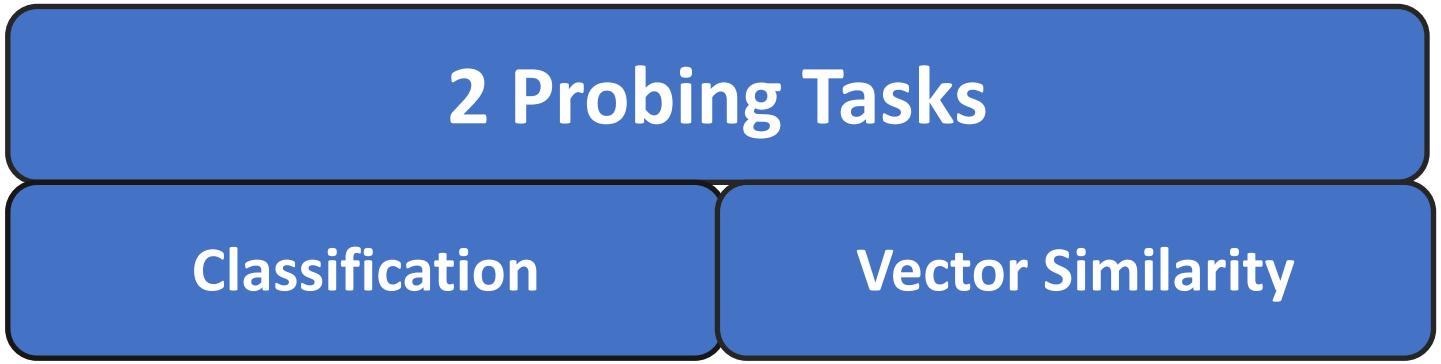
Dataset

- To evaluate we create a custom corpus
  - 1000 idioms curated from SLIDE dataset
    - Selected for min length > 3 and variation in concreteness, abstractness etc.
  - 2000 idioms in context of regular language usage(Reddit Comments)with paraphrases and non paraphrases (2 of each per sample) created by our team.

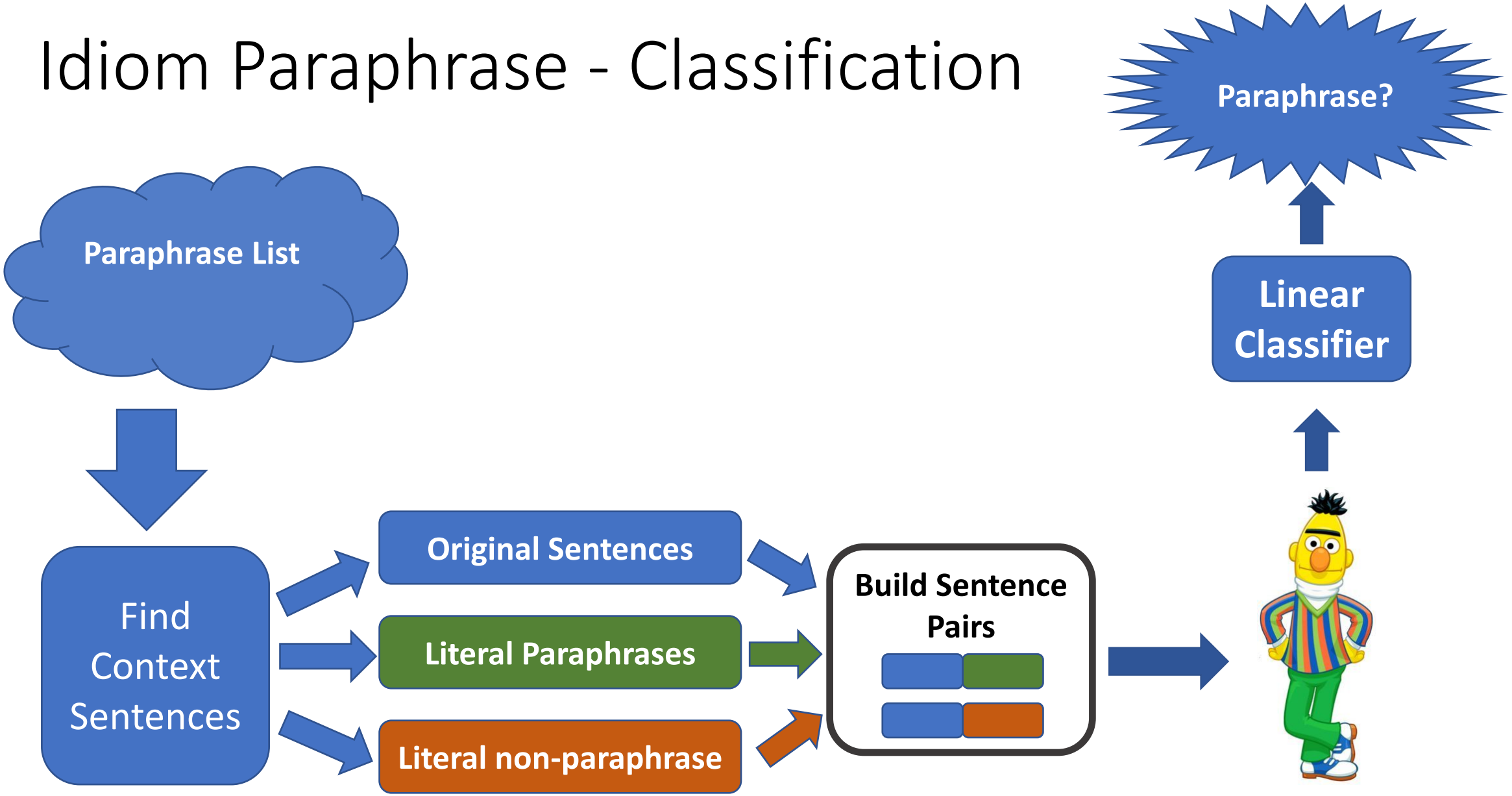
Idiom Paraphrase Evaluation

Given Sentence 1 and Sentence 2:

Can a pretrained language model tell us if Sentence 1 is a Paraphrase of Sentence 2?



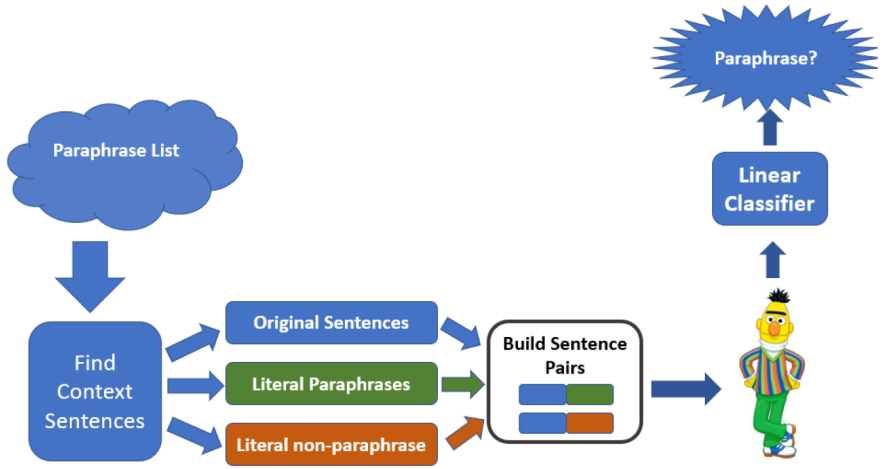
Idiom Paraphrase - Classification



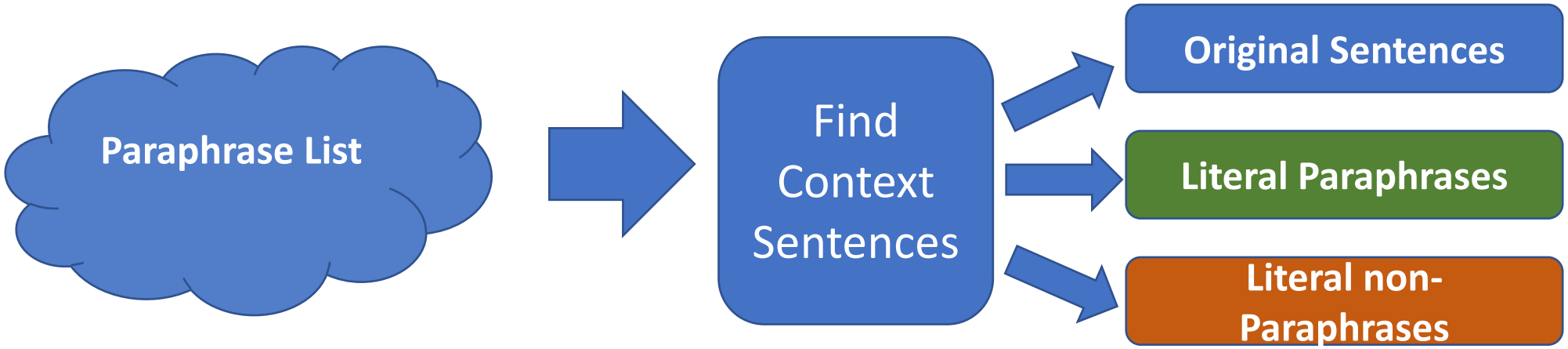
https://www.clipart.email/

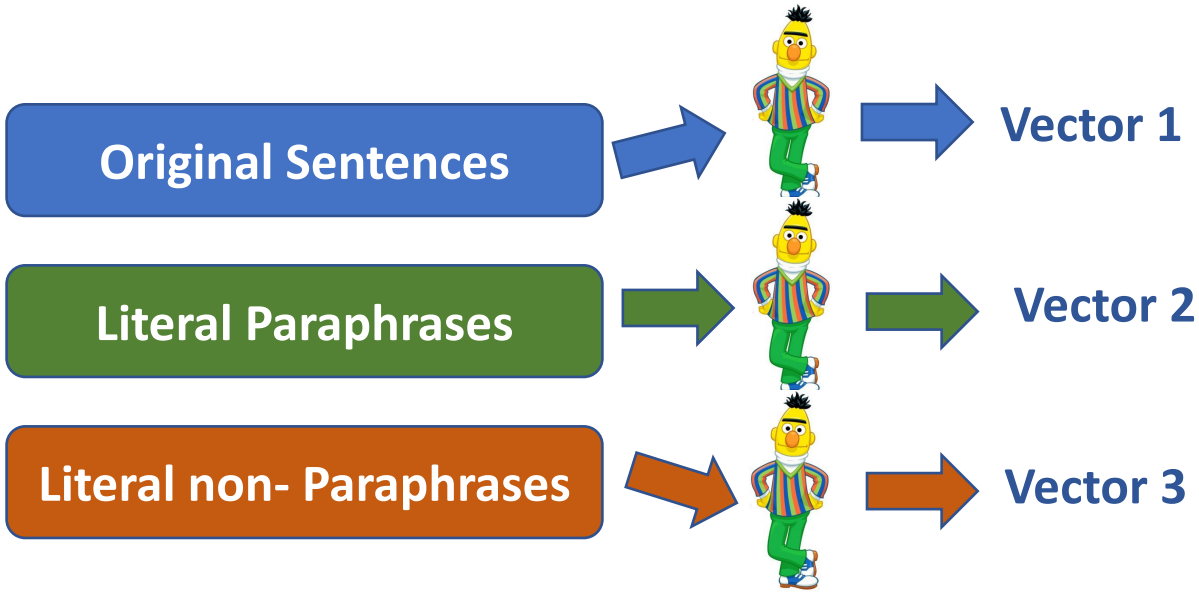
Classification - Variations

2 areas of variation with 2 options each	
Which embeddings to use?	**Which Bert to Use?**
• Feed s1 through BERT • Feed s2 through BERT • Combine $embedding_{s1}$ and $embedding_{s2}$	BERT Base with no fine-tuning
• Feed s1 + s2 through BERT • Use CLS token	BERT finetuned on paraphrase detection (but not on idioms)



Idiom Paraphrase – Vector Similarity

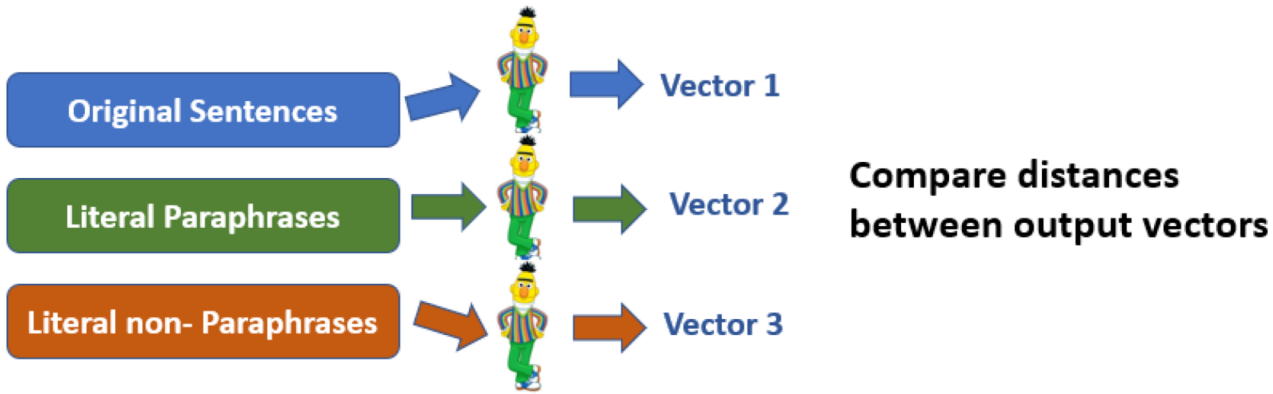




Compare distances between output vectors

Vector Similarity – Details

- Choice of non-paraphrases is important – lexical overlap



She let the cat out of the bag

She revealed the secret

The cat jumped out of the bag

Vector Similarity – Variations

- Which vectors to look at?
  - Representation of full sentence, or select words?
    - “cat” vs. “secret”
  - Which layer(s) of BERT?

- What comparison measure to use?

Summary – Idiom Paraphrase Experiment

- Two Tasks:
  - Classifier: Given two sentences, where one contains an idiom, classify as paraphrase or not paraphrase.
  - Vector Similarity: Gather BERT representations for variations on sentences with idioms (true paraphrases, false paraphrases). Compare the distances between these vectors.
- Look for trends in results – does vector similarity relate to classifier success?

Questions?

Thanks!

Group 1:

Josh Tanner, Paige Finkelstein, Wes Rose,
Elena Khasanova, and Daniel Campos