第12回最先端NLP勉強会（2020.9.26）

悩みの種

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

論文: https://www.aclweb.org/anthology/Q19-1027.pdf

発表動画: https://vimeo.com/422205958

著者スライド:

https://vered1986.github.io/papers/Lexical_Composition_TACL_EMNLP_2019_Presentation.pdf

紹介者: 笹野遼平（名大）

論文の概要

- トピック: Lexical Composition（語の複合）
  1. Detecting **meaning shift** [MS]
     - *carry on* ≠ *carry* + *on*, *guilt trip* ≠ *guilt* + *trip*
  2. Recovering **implicit meaning** [IM]
     - *olive oil*: made of olives ⇔ *baby oil*: made for babies

- 各種意味表現が上記の現象を扱えるかを検証
  - ✓ 文脈化単語表現(ELMo, GPT, BERT)は静的な単語埋め込み(word2vec, GloVe, fastText)よりうまく扱える
  - ✓ implicit meaningの復元精度はいまだに人間の精度との隔たりが大きい

6 Representations × 6 Tasks

- 6 Representations:

	training objective	corpus (#words)	output dimension	basic unit
word embeddings				
WORD2VEC	Predicting surrounding words	Google News (100B)	300	word
GLOVE	Predicting co-occurrence probability	Wikipedia + Gigaword 5 (6B)	300	word
FASTTEXT	Predicting surrounding words	Wikipedia + UMBC + statmt.org (16B)	300	subword
contextualized word embeddings				
ELMO	Language model	1B Word Benchmark (1B)	1024	character
OPENAI GPT	Language model	BooksCorpus (800M)	768	subword
BERT	Masked language model (Cloze)	BooksCorpus + Wikipedia (3.3B)	768	subword

- 6 Composition Tasks:
  - 既存データを活用、ただし、基本的に分類問題の形式に変換

1. Verb-Particle Constructions (VPC) Classification [MS]
2. Light Verb Constructions (LVC) Classification [MS]
3. Noun Compound (NC) Literality [MS]
4. Noun Compound (NC) Relations [IM]
5. Adjective-Noun (AN) Attributes [IM]
6. Identifying Phrase Type [MS & IM] （これだけ系列ラベリング）

共通かつシンプルなモデルで検証

- 解析対象の範囲の最初と 最後のベクトルをconcatしたものを入力し分類 (paraphrase等もあれば入力: u')
  - $\vec{x} = [\vec{u}_i; \vec{u}_{i+k}; \vec{u'}_1; \vec{u'}_l]$
  - $\vec{o} = \mathrm{softmax}(W \cdot \mathrm{ReLU}(\mathrm{Dropout}(h(\vec{x}))))$
- 使用するLayer: Top or All (=タスクごとに重みを学習し足し合せ)
- 各unitのembeddingを３つの方法でencode
  1. biLM: biLSTMに通す $\boxed{\vec{u}_1, \ldots, \vec{u}_n = \mathrm{biLSTM}(\vec{v}_1, \ldots, \vec{v}_n)}$
  2. Att: Self-attention $\boxed{\vec{u}_i = [\vec{v}_i; \sum_{j=1}^{n} a_{i,j} \cdot \vec{v}_j],\ \vec{a}_i = \mathrm{softmax}(\vec{v}_i^T \cdot \vec{v})}$
  3. None: そのまま $\boxed{\vec{u}_1, \ldots, \vec{u}_n = \vec{v}_1, \ldots, \vec{v}_n}$

比較対象

- Human Performances
  - テストセットごとに100事例を再アノテーション
  - AMTで受理率98%以上、500以上のhuman intelligence tasksの実績があり、品質試験を通過したworkerのみ
  - 3 workersのmajority labelを採用
- Majority Baselines
  - 訓練とテストで分布が異なるので２値分類でも50%以下になりうる（train, val, testで語彙が重複しないようにsplit）
  - VPCの場合:
    - Train: V={take, get}, #true=710, #false=209
    - Val: V={make}, #true=116, #false=93
    - Test: V={do, have, give}, #true=52, #false=168 (⇒ 52/220=23.6%)

実験結果

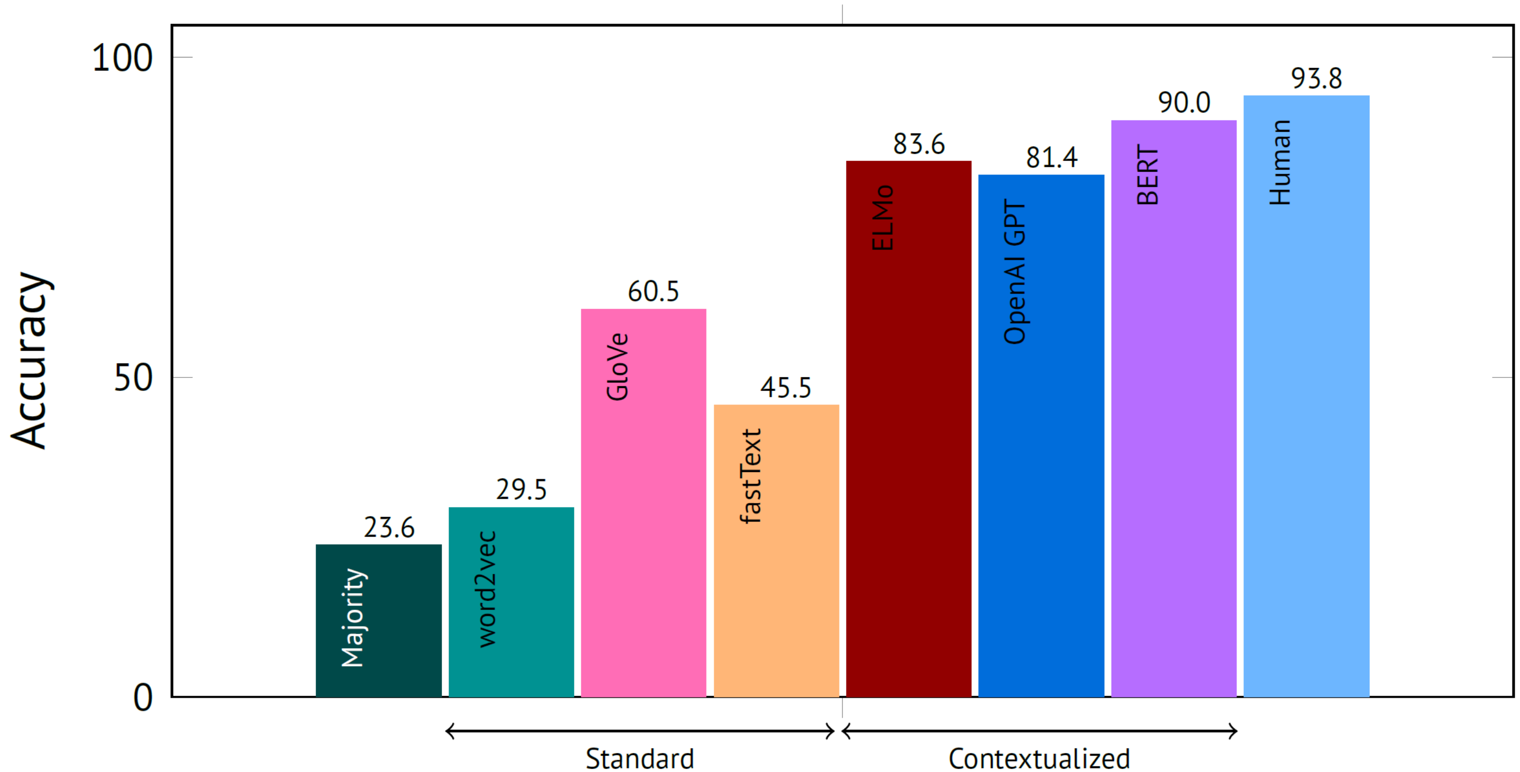
Verb-Particle Constructions [MS]

VPC

We did get on together

Non-VPC

Which response did you get on that?

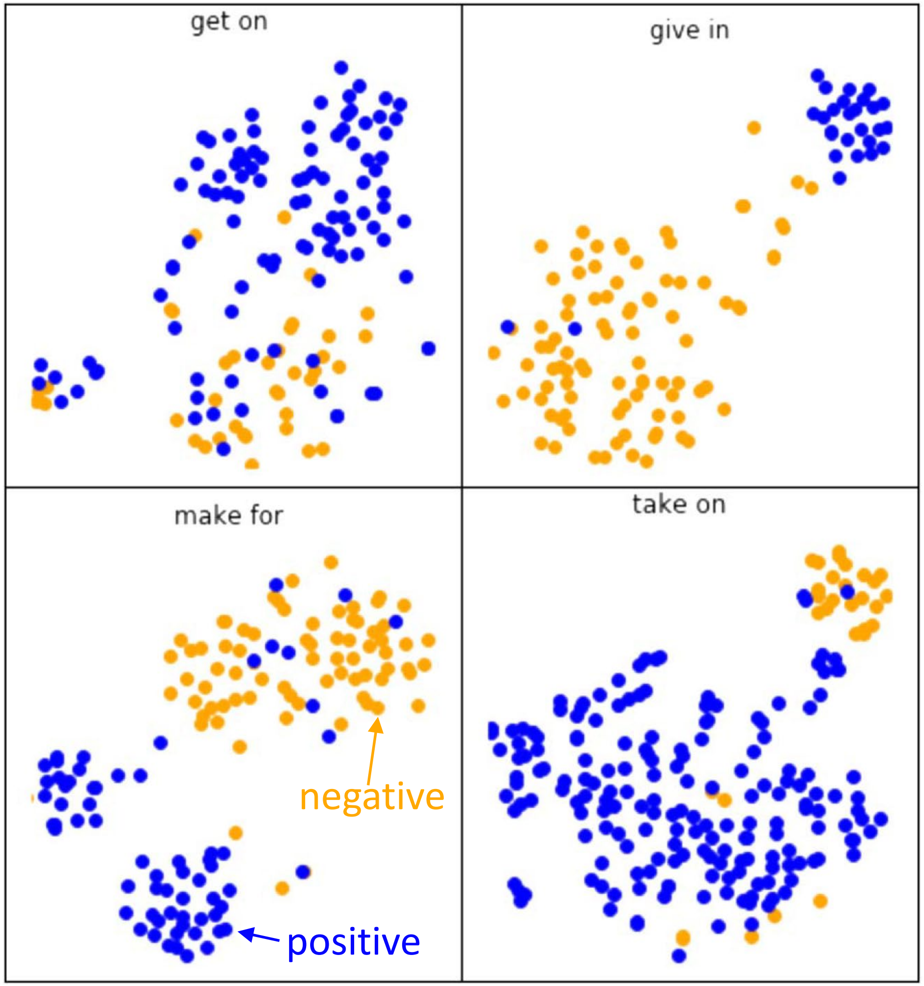


VPCを本当に捉えているか？

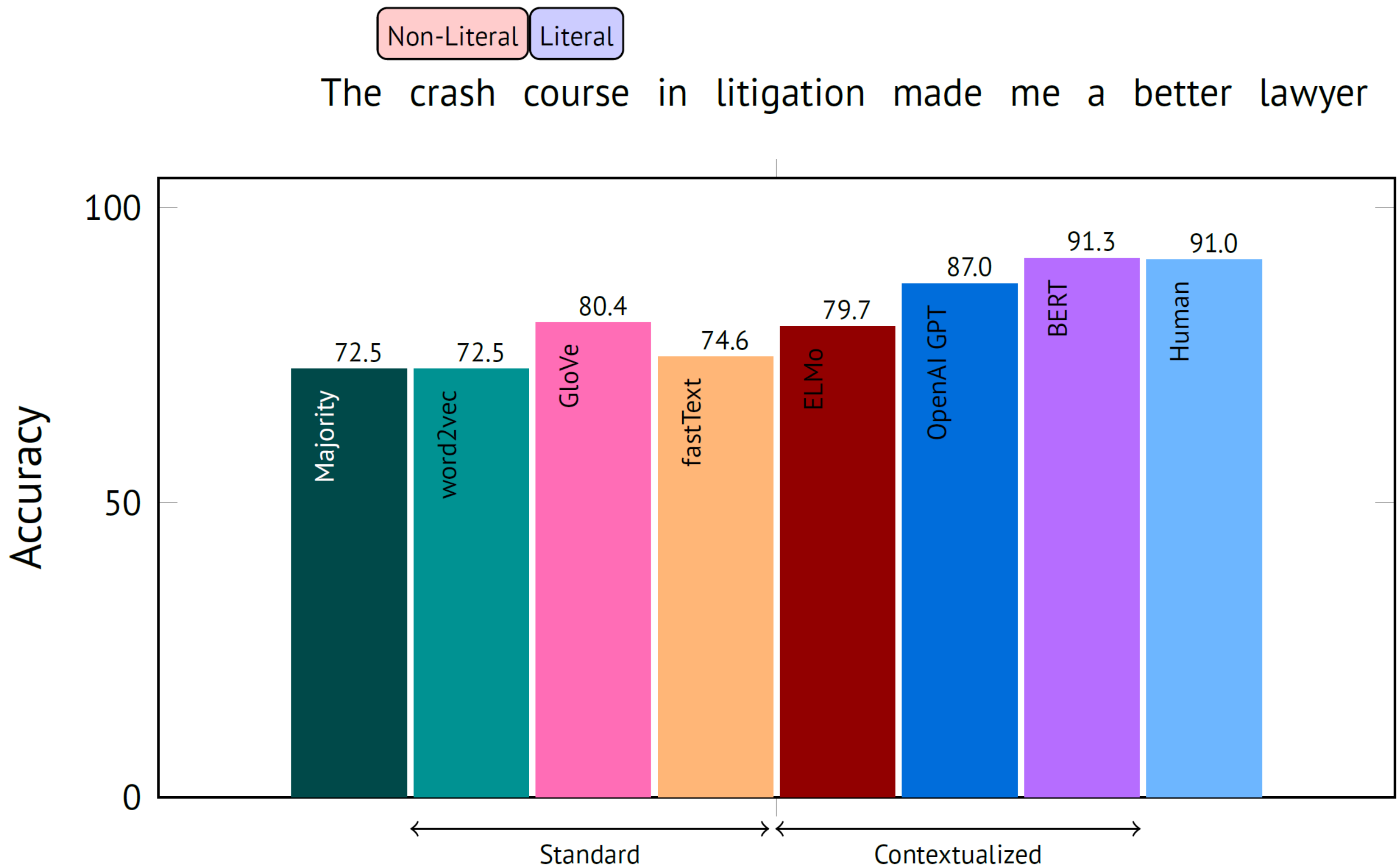
- 曖昧な表現に注目
  - VPC, non-VPCがともに８事例以上ある表現
- 構成要素のBERTベクトルをconcatしたものをt-SNEで２次元投影

⇒BERTはVPCとnon-VPCの違いを捉えている

- 実際にはどちらがVPCかも捉えているらしい



Noun Compound Literality [MS]



妥当な置換語を予測できるか？

ELMo	OpenAI GPT	BERT
The Queen and her husband were on a train **trip** from Sydney to Orange.		
ride	to	travelling
carriage	headed	running
journey	heading	journey
heading	that	going
carrying	and	headed

Creating a guilt **trip** in another person may be considered to be psychological manipulation...		
tolerance	that	reaction
fest	so	feeling
avoidance	trip	attachment
onus	he	sensation
association	she	note

- literalな事例の置き換えはわりと上手くいっている（特にBERT）
- non-literalについては上手くできる例は限定的（他の例は論文参照）
- それがnon-literalであることは捉えているらしい

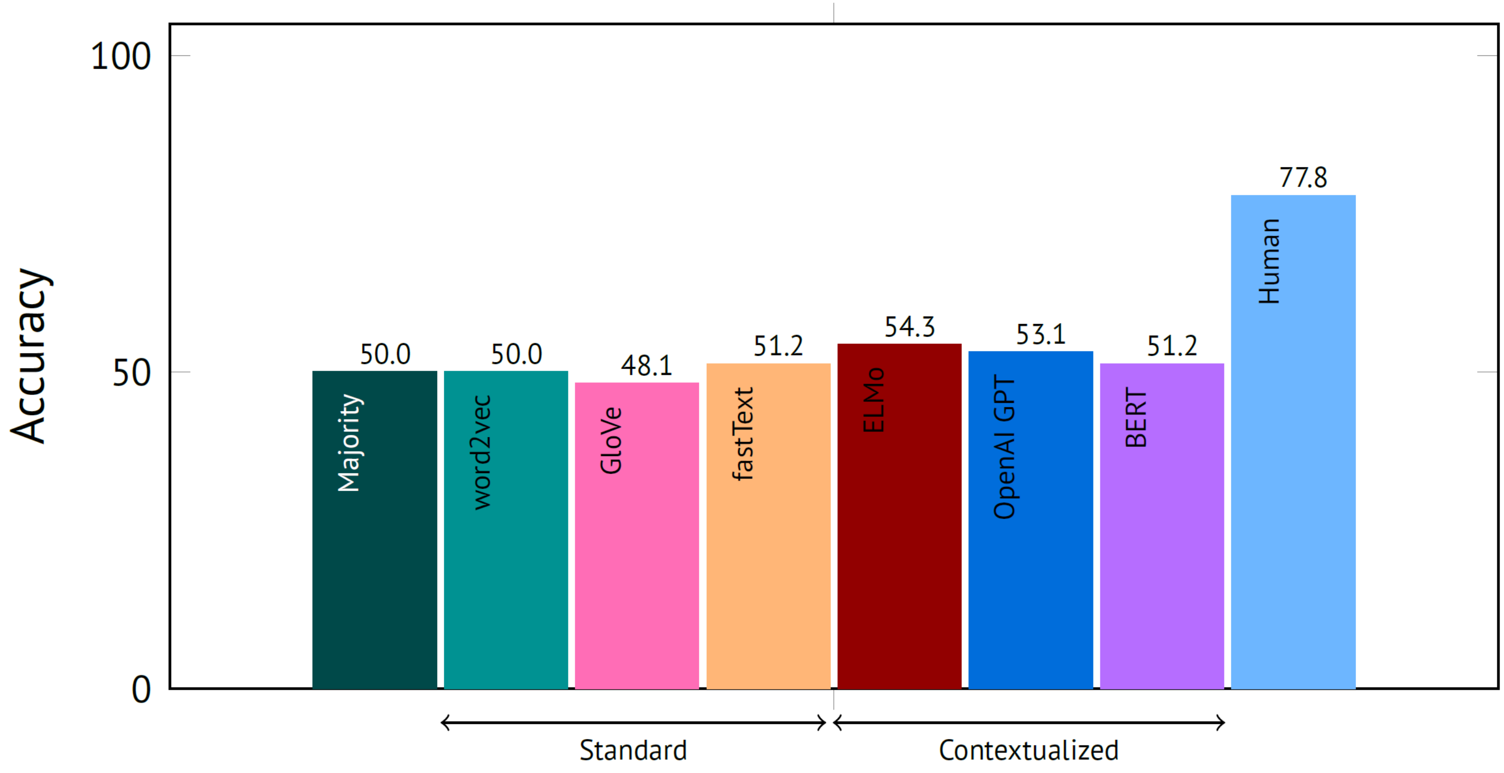
Noun Compound Relations [IM]

Paraphrase

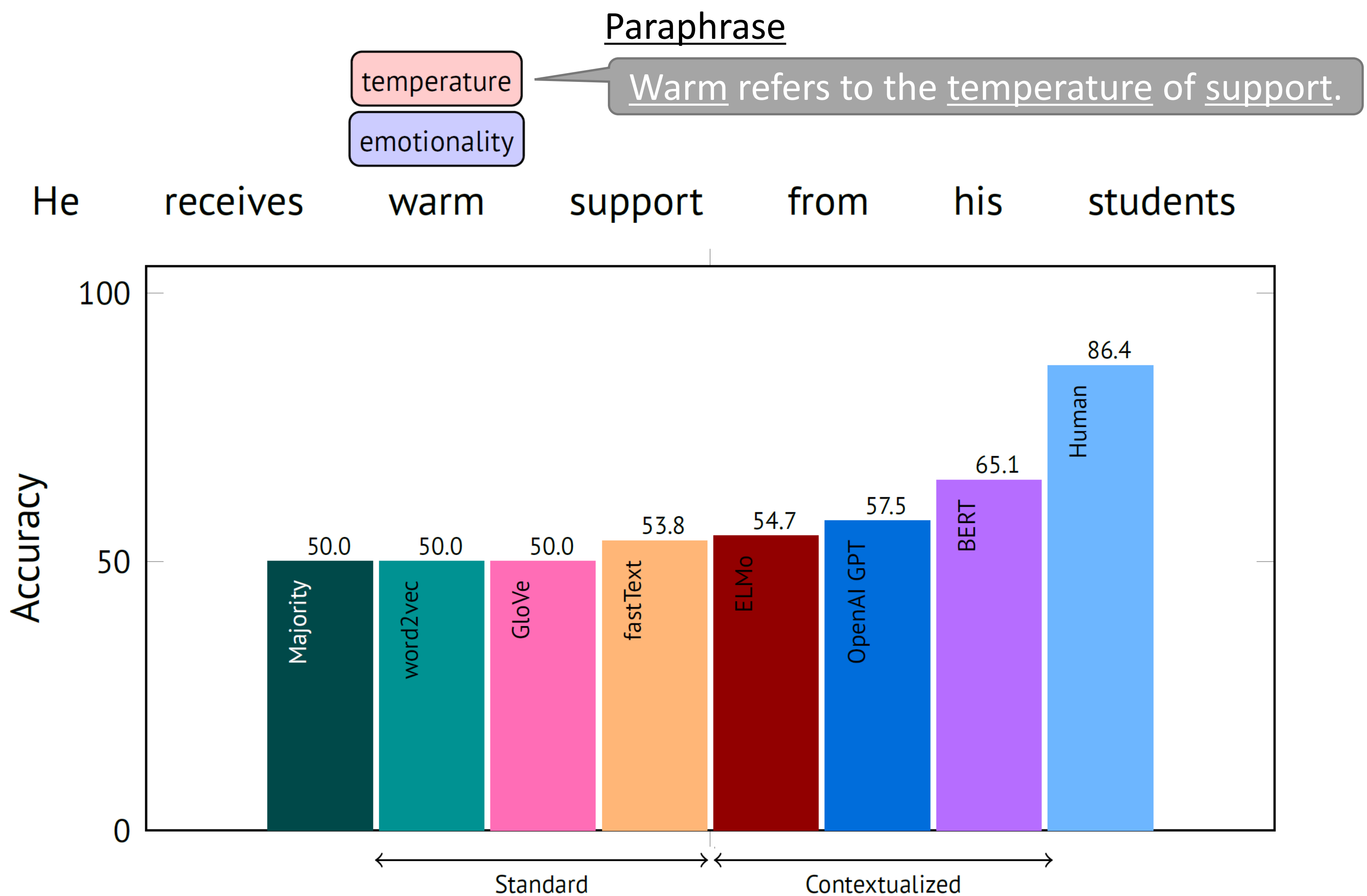
Road forecasted for access season

Road that makes access possible

The township is served by three access roads .



Adjective-Noun Attributes [IM]



層とエンコード手法の選択について

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**

- 各層の重み付き和を使う方が良い傾向
  - 実際にはtopとbottom層の混合が学習されていることが多いらしい
- エンコード手法ついては顕著な優劣なし
  - meaning shiftタスクについてはAttとNoneが優勢
  - implicit meaningタスクについてはbiLMが優勢

今後の方向性: 人間と同じように フレーズの意味を捉える

- L2学習者がどうidiomを処理するか？ [Cooper'99]

1. Infer from context: 28% (57% success rate)
   - より“拡張された”文脈(stories)を利用
   - e.g., Characters in the story, Relationships between them, …
2. Rely on literal meaning: 19% (22% success rate)
   - *“Robert knew he was robbing the cradle by dating a 16-year-old girl”*
   - Knowledge + Reasoning:
     - Cradle is something you put the baby in

     ⇒ Stealing a child from a mother”

     ⇒ *“rob the cradle”* means having relations with a very younger person

おわりに

- 論文の主な貢献
  1. 既存データセットを活用し統一的に各種意味表現のフレーズ処理性能を分析
  2. フレーズ分析のためのフレームワークを構築（データ等も公開）

- 結果に関して
  - 文脈化埋め込みの方が良さそうなのは概要に“as expected”と書かれているとおり予想の範囲内
  - Meaning shiftにおけるBERTの精度が高さはわりと不思議
    （違いを捉えているだけでなくどちらがshiftしたものかも捉えている！）
  - Implicit meaningの復元に関してはモデルが適切でない可能性
  - Cross-validationは行っていないので結果の一般性はやや疑問
  - 人間の精度の妥当性もやや気になる（学習データを見ていない, “I can’t tell”, “the sentence does not make sense”という選択肢の存在, そもそも人間の精度とは？）

補足

Composition Tasks

Task	Data Source	Train/val/test Size	Input	Output
VPC Classification	Tu and Roth (2012)	919/209/220	sentence s VP = w_1 w_2	is VP a VPC?
LVC Classification	Tu and Roth (2011)	1521/258/383	sentence s span = w_1 ... w_k	is the span an LVC?
NC Literality	Reddy et al. (2011) Tratz (2011)	2529/323/138	sentence s NC = w_1 w_2 target w $\in \{w_1, w_2\}$	is w literal in NC?
NC Relations	SemEval 2013 Task 4 (Hendrickx et al., 2013)	1274/162/130	sentence s NC = w_1 w_2 paraphrase p	does p explicate NC?
AN Attributes	HeiPLAS (Hartung, 2015)	837/108/106	sentence s AN = w_1 w_2 paraphrase p	does p describe the attribute in AN?
Phrase Type	STREUSLE (Schneider and Smith, 2015)	3017/372/376	sentence s	label per token

Worker Agreement

Task	**Agreement**	**Example Question**
VPC Classification	84.17%	*I feel there are others far more suited to* ***take on*** *the responsibility.* What is the verb in the highlighted span? (take/take on)
LVC Classification	83.78%	*Jamie* ***made a decision*** *to drop out of college.* Mark all that apply to the highlighted span in the given context: 1. It describes an action of ''*making something*'', in the common meaning of ''*make*''. 2. The essence of the action is described by ''*decision*''. 3. The span could be rephrased without ''*make*'' but with a verb like ''*decide*'', without changing the meaning of the sentence.
NC Literality	80.81%	*He is driving down memory* ***lane*** *and reminiscing about his first love.* Is ''lane'' used literally or non-literally? (literal/non literal)
NC Relations	86.21%	*Strawberry shortcakes were held as celebrations of the* ***summer fruit*** *harvest.* Can ''summer fruit'' be described by ''fruit that is ripe in the summer''? (yes/no)
AN Attributes	86.42%	*Send my* ***warm*** *regards to your parents.* Does ''warm'' refer to temperature? (yes/no)