Multi-word Units Under the Magnifying Glass

Vered Shwartz
Natural Language Processing Lab, Bar-Ilan University

Talk @ ONLP, December 26, 2018

Multi-Word Units (MWUs)*

- A sequence of consecutive words that creates a new concept

Multi-Word Units (MWUs)*

- A sequence of consecutive words that creates a new concept
  - **Noun compounds:** *flea market*, *flea bite*, *flea bite treatment*, ...
  - **Adjective-noun compositions:** *hot tea*, *hot day*, ...
  - **Verb-particle constructions:** *wake up*, *let go*, ...
  - **Light-verb constructions:** *make a decision*, *take a walk*, ...
  - **Idioms:** *look what the cat dragged in*, ...

Multi-Word Units (MWUs)*

- A sequence of consecutive words that creates a new concept
  - **Noun compounds:** *flea market*, *flea bite*, *flea bite treatment*, ...
  - **Adjective-noun compositions:** *hot tea*, *hot day*, ...
  - **Verb-particle constructions:** *wake up*, *let go*, ...
  - **Light-verb constructions:** *make a decision*, *take a walk*, ...
  - **Idioms:** *look what the cat dragged in*, ...

- May combine in a non-trivial way
  - Implicit meaning
  - Non-literal word usage

Multi-Word Units (MWUs)*

- A sequence of consecutive words that creates a new concept
  - **Noun compounds:** *flea market*, *flea bite*, *flea bite treatment*, ...
  - **Adjective-noun compositions:** *hot tea*, *hot day*, ...
  - **Verb-particle constructions:** *wake up*, *let go*, ...
  - **Light-verb constructions:** *make a decision*, *take a walk*, ...
  - **Idioms:** *look what the cat dragged in*, ...

- May combine in a non-trivial way
  - Implicit meaning
  - Non-literal word usage

* Also referred to as Multi-Word Expressions or phrases

Previous MWUs Representations

- **Compositional Distributional Representations:**
  - $vec(\textit{olive oil}) = f(vec(\textit{olive}), vec(\textit{oil}))$
  - Many ways to learn f [Mitchell and Lapata, 2010, Zanzotto et al., 2010, Dinu et al., 2013]
  - Usually applied to AN or NC, limited to specific number of words

Previous MWUs Representations

- **Compositional Distributional Representations:**
  - $vec(olive\ oil) = f(vec(olive), vec(oil))$
  - Many ways to learn f [Mitchell and Lapata, 2010, Zanzotto et al., 2010, Dinu et al., 2013]
  - Usually applied to AN or NC, limited to specific number of words

- **Phrase Embeddings:**
  - Arbitrarily long phrases

Previous MWUs Representations

- **Compositional Distributional Representations:**
  - $vec(olive\ oil) = f(vec(olive), vec(oil))$
  - Many ways to learn f [Mitchell and Lapata, 2010, Zanzotto et al., 2010, Dinu et al., 2013]
  - Usually applied to AN or NC, limited to specific number of words

- **Phrase Embeddings:**
  - Arbitrarily long phrases
  - Supervision from PPDB [Wieting et al., 2015]
    - Limited in coverage

Previous MWUs Representations

- **Compositional Distributional Representations:**
  - $vec(\textit{olive oil}) = f(vec(\textit{olive}), vec(\textit{oil}))$
  - Many ways to learn f [Mitchell and Lapata, 2010, Zanzotto et al., 2010, Dinu et al., 2013]
  - Usually applied to AN or NC, limited to specific number of words

- **Phrase Embeddings:**
  - Arbitrarily long phrases
  - Supervision from PPDB [Wieting et al., 2015]
    - Limited in coverage
  - Generalizing word2vec [Poliak et al., 2017]
    - Can compose vectors for unseen phrases
    - Naive composition, doesn't handle the complexity of phrases

Enter contextualized word embeddings!

- Represent a word *in context*
  - Good for word sense induction
- Trained as language models
  - On a large corpus
  - Capture world knowledge
- Improve performance of various NLP applications
- Named after characters from Sesame Street

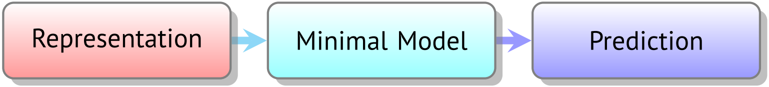
Enter contextualized word embeddings!

- Represent a word *in context*
  - Good for word sense induction
- Trained as language models
  - On a large corpus
  - Capture world knowledge
- Improve performance of various NLP applications
- Named after characters from Sesame Street

Are meaningful MWU representations built-in in these models?

Probing Tasks

- Simple tasks designed to test a single linguistic property [Adi et al., 2017, Conneau et al., 2018]



Probing Tasks

- Simple tasks designed to test a single linguistic property [Adi et al., 2017, Conneau et al., 2018]

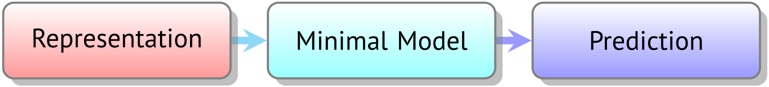
Representation → Minimal Model → Prediction

SkipThoughts(*s*)
InferSent(*s*)
...

What is *s*'s length?
Is *w* in *s*?
...

Probing Tasks

- Simple tasks designed to test a single linguistic property [Adi et al., 2017, Conneau et al., 2018]



SkipThoughts(*s*)
InferSent(*s*)
...

What is *s*'s length?
Is *w* in *s*?
...

- We follow the same for MWUs, with various representations

Representations

Word Embeddings	Sentence Embeddings	Contextualized Word Embeddings
word2vec	SkipThoughts	ELMo
GloVe	InferSent*	OpenAI Transformer
fastText	GenSen*	BERT

* supervised

Tasks and Results

1. MWU Type
Task Definition

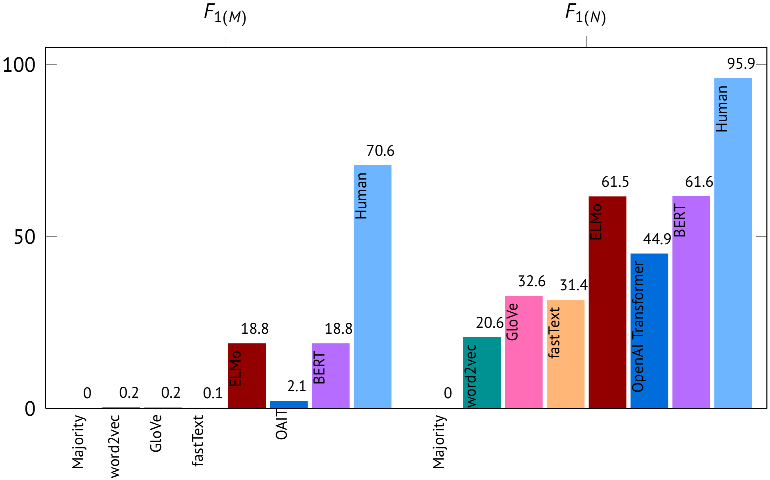
- **Dataset:** Wiki50 corpus [Vincze et al., 2011]
- **Input:** sentence
- **Goal:** sequence labeling to BIO tags
  - **MWUs**: noun compounds, adjective-noun compositions, idioms, light verb constructions, verb-particle constructions
  - **Named entities**: person, location, organization

- **Example:**

O	B-MW_VPC	I-MW_VPC	B-MW_NC	I-MW_NC	O	O	O	O
Authorities	meted	out	summary	justice	in	cases	as	this

1. MWU Type

Results



(1) Identifying MWU type is difficult; (2) Named entities are easier; (3) Context helps

2. Noun Compound Literality

A constituent word may be used in a non-literal way

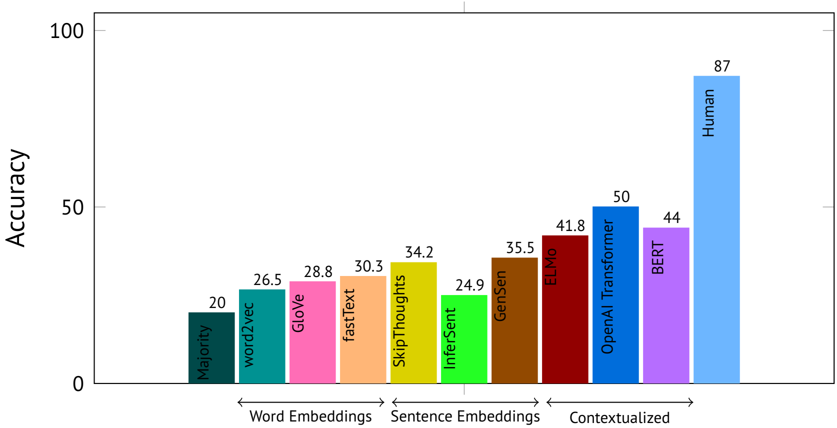
2. Noun Compound Literality
Task Definition

- **Dataset:** based on [Reddy et al., 2011] and [Tratz, 2011]
- **Input:** sentence s, target word $w \in s$ (part of NC)
- **Goal:** is w literal in NC?

- **Example:**

	Non-Literal	Literal							
The	crash	course	in	litigation	made	me	a	better	lawyer

2. Noun Compound Literality
Results



(1) word embeddings < sentence embeddings < contextualized; (2) Far from humans

2. Noun Compound Literality
Analysis

ELMo	OpenAI Transformer	BERT
A search team located the [crash]$_L$ site and found small amounts of human remains.		
landfill	body	archaeological
wreckage	place	burial
Web	man	wreck
crash	missing	excavation
burial	location	grave
After a [crash]$_N$ course in tactics and maneuvers, the squadron was off to the war...		
crash	few	short
changing	while	successful
collision	moment	rigorous
training	long	brief
reversed	couple	training

(1) BERT > ELMo, both reasonable
(2) OpenAI Transformer errs due to uni-directionality

2. Noun Compound Literality
Analysis

ELMo	OpenAI Transformer	BERT
The gold/[silver]$_L$ price ratio is often analyzed by traders, investors, and buyers.		
silver	platinum	silver
blue	black	copper
platinum	gold	platinum
purple	silver	gold
yellow	red	diamond
Growing up with a [silver]$_N$ spoon in his mouth, he was always cheerful...		
silver	mother	wooden
rubber	father	greasy
iron	lot	big
tin	big	silver
wooden	man	little

Things get tougher when both constituent nouns are non-literal!

3. Noun Compound Relations

- NCs express semantic relations between the constituent words

3. Noun Compound Relations

- NCs express semantic relations between the constituent words
- May require world knowledge and common sense to interpret

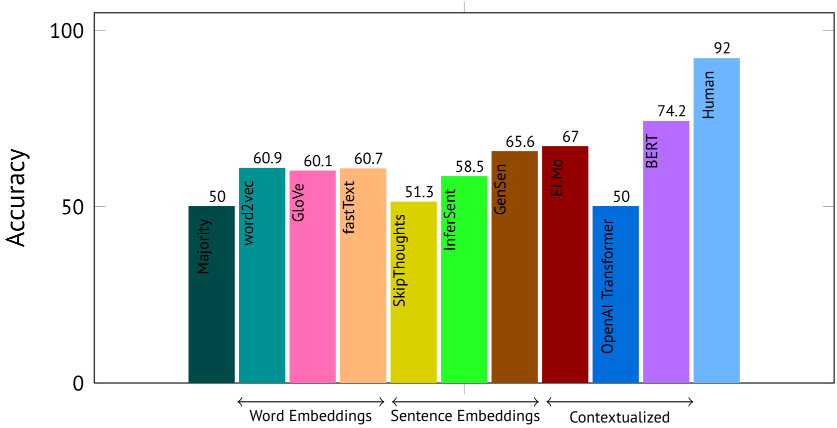
3. Noun Compound Relations
Task Definition

- **Dataset:** based on [Hendrickx et al., 2013]
- **Input:** sentence s, NC $\in$ s, paraphrase p
- **Goal:** does p explicate NC?

- **Example:** *access road*

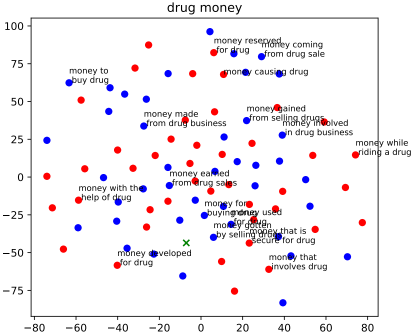
  Road that makes access possible ✓
  Road forecasted for access season ×

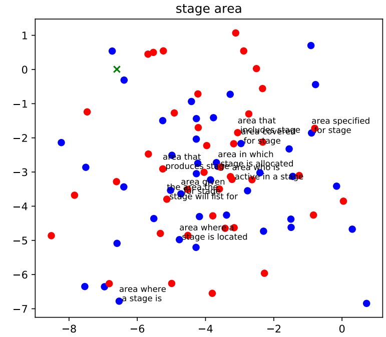
3. Noun Compound Relations
Results



(1) word embeddings < sentence embeddings < contextualized; (2) Far from humans; (3) Open AI Transformer fails

3. Noun Compound Relations
Analysis





No clear signal from BERT. Capturing implicit information is challenging!

4. Adjective-Noun Relations

Adjectives select different attributes of the noun they combine with

The hot debate about the hot office (or: the cold war over the cold office)

4. Adjective-Noun Relations
Task Definition

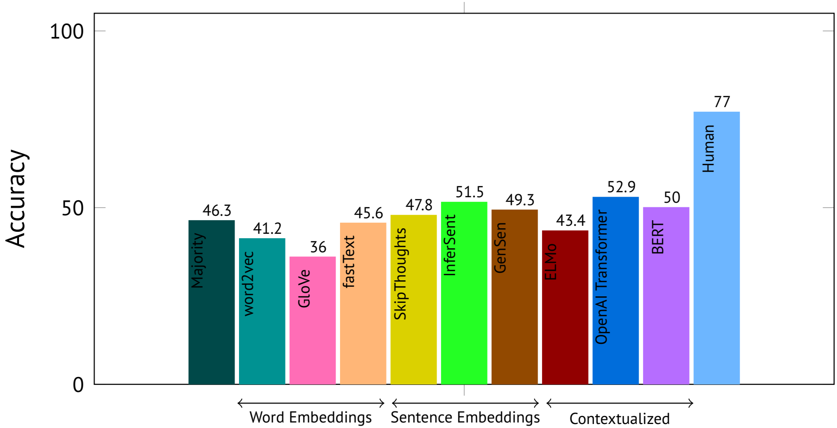
- **Dataset:** based on [Hartung, 2015]
- **Input:** sentence s, AN $\in$ s, attribute w
- **Goal:** is the attribute w conveyed in AN?

- **Example:** *warm support*:

  `temperature` $\times$

  `emotionality` $\checkmark$

4. Adjective-Noun Relations
Results



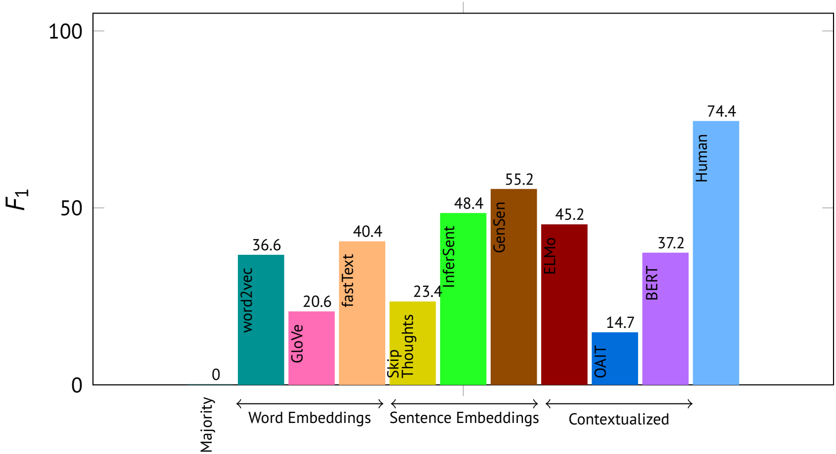
Best model performs only slightly better than majority
(Capturing implicit information is challenging)

5. Adjective-Noun Entailment
Task Definition

- **Dataset:** [Pavlick and Callison-Burch, 2016]
- **Input:** premise p, hypothesis h, differ by a single adjective
- **Goal:** p $\rightarrow$ h?

- **Example:**
  p: *Most people die in the class to which they were born.*
  h: *Most people die in the* ***social*** *class to which they were born.* ✓

5. Adjective-Noun Entailment
Results



Bad performance for all models, best for sentence embeddings trained on RTE

6. Verb-Particle Classification

VPC meanings differ from their verbs' meanings

6. Verb-Particle Classification
Task Definition

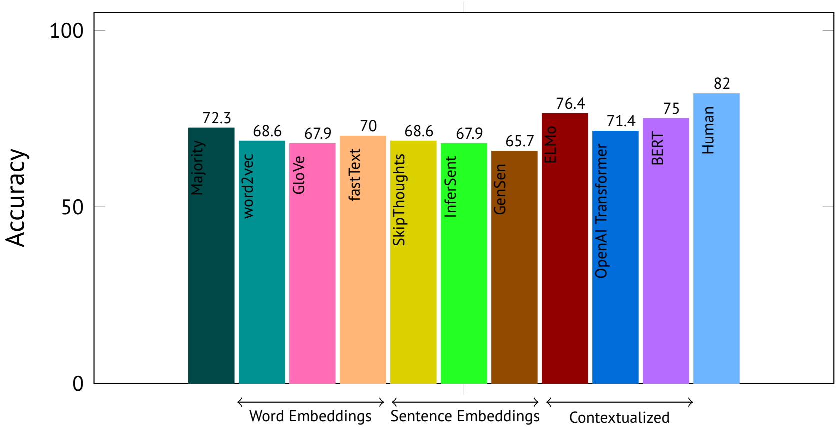
- **Dataset:** [Tu and Roth, 2012]
- **Input:** sentence s, VP $\in$ s
- **Goal:** is VP a VPC?

- **Example:**

We did get on together (VPC: "get on")

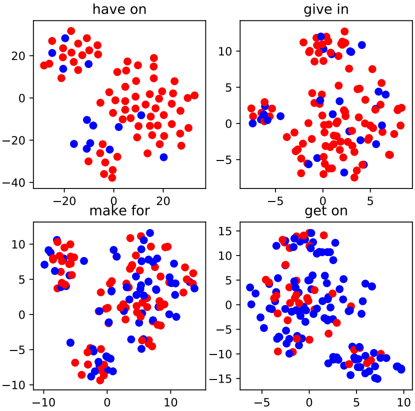
Which response did you get on that? (Non-VPC: "get on")

6. Verb-Particle Classification
Results



Similar performance for all models. Is the good performance merely due to label imbalance?

6. Verb-Particle Classification
Analysis



Very weak signal from ELMo. Mostly performs well due to label imbalance.

Future Directions

Can we learn MWUs like humans do?

- [Cooper, 1999]: how do L2 learners process idioms?
  - **Infer from context:** 28% (57% success rate)
  - **Rely on literal meaning:** 19% (22% success rate)
  - ...

Inferring from context

We need richer context modeling

Furious Meghan Markle says she won't fall for dad's 'crocodile tears' after he claimed 'she'd be better off if he were dead'

FURIOUS Meghan Markle has said she won't fall for her dad's "crocodile tears" after he claimed "she'd be better off if he were dead".

The Duchess of Sussex reportedly told pals Thomas Markle is using "emotional blackmail" to try and manipulate her but she's had "enough already".

Previous news stories may help understand that "crocodile tears" refer to manipulative behavior

[Asl, 2013]: L2 learners interpret idioms with more success through extended contexts (stories) than through sentential contexts

Relying on literal meaning

We need world knowledge

“Cradle is something that you put the baby in”

“You’re stealing a child from a mother”

*“So **robbing the cradle** is like dating a really young person”*

[Cooper, 1999]

Recap

1. **Testing Existing Pre-trained Representations**

   Contextualized word embeddings provide better MWU representations, but there is still a long way to go

Recap

1. **Testing Existing Pre-trained Representations**

   Contextualized word embeddings provide better MWU representations, but there is still a long way to go

2. **Future Directions**

   To represent MWUs like humans do, we need better context and world knowledge modeling

Recap

1. **Testing Existing Pre-trained Representations**

   Contextualized word embeddings provide better MWU representations, but there is still a long way to go

2. **Future Directions**

   To represent MWUs like humans do, we need better context and world knowledge modeling

Thank you!

References I

[Adi et al., 2017] Adi, Y., Kermany, E., Belinkov, Y., Lavi, O., and Goldberg, Y. (2017). Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *Proceedings of ICLR Conference Track*.

[Asl, 2013] Asl, F. M. (2013). The impact of context on learning idioms in efl classes. *TESOL Journal*, 37(1):2.

[Conneau et al., 2018] Conneau, A., Kruszewski, G., Lample, G., Barrault, L., and Baroni, M. (2018). What you can cram into a single vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136. Association for Computational Linguistics.

[Cooper, 1999] Cooper, T. C. (1999). Processing of idioms by l2 learners of english. *TESOL quarterly*, 33(2):233–262.

[Dinu et al., 2013] Dinu, G., Pham, N. T., and Baroni, M. (2013). General estimation and evaluation of compositional distributional semantic models. In *Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality*, pages 50–58, Sofia, Bulgaria. Association for Computational Linguistics.

[Hartung, 2015] Hartung, M. (2015). *Distributional Semantic Models of Attribute Meaning in Adjectives and Nouns*. Ph.D. thesis, Heidelberg University.

[Hendrickx et al., 2013] Hendrickx, I., Kozareva, Z., Nakov, P., Ó Séaghdha, D., Szpakowicz, S., and Veale, T. (2013). Semeval-2013 task 4: Free paraphrases of noun compounds. In *SemEval*, pages 138–143.

[Mitchell and Lapata, 2010] Mitchell, J. and Lapata, M. (2010). Composition in distributional models of semantics. *Cognitive science*, 34(8):1388–1429.

[Pavlick and Callison-Burch, 2016] Pavlick, E. and Callison-Burch, C. (2016). Most "babies" are "little" and most "problems" are "huge": Compositional entailment in adjective-nouns. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2164–2173, Berlin, Germany. Association for Computational Linguistics.

[Poliak et al., 2017] Poliak, A., Rastogi, P., Martin, M. P., and Van Durme, B. (2017). Efficient, compositional, order-sensitive n-gram embeddings. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 503–508, Valencia, Spain. Association for Computational Linguistics.

References II

[Reddy et al., 2011] Reddy, S., McCarthy, D., and Manandhar, S. (2011). An empirical study on compositionality in compound nouns. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 210–218, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

[Tratz, 2011] Tratz, S. (2011). *Semantically-enriched parsing for natural language understanding*. University of Southern California.

[Tu and Roth, 2012] Tu, Y. and Roth, D. (2012). Sorting out the most confusing english phrasal verbs. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 65–69, Montréal, Canada. Association for Computational Linguistics.

[Vincze et al., 2011] Vincze, V., Nagy T., I., and Berend, G. (2011). Multiword expressions and named entities in the wiki50 corpus. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 289–295. Association for Computational Linguistics.

[Wieting et al., 2015] Wieting, J., Bansal, M., Gimpel, K., and Livescu, K. (2015). Towards universal paraphrastic sentence embeddings. *CoRR*, abs/1511.08198.

[Zanzotto et al., 2010] Zanzotto, F. M., Korkontzelos, I., Fallucchi, F., and Manandhar, S. (2010). Estimating linear models for compositional distributional semantics. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1263–1271. Association for Computational Linguistics.