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:**
  - $\text{vec}(\text{olive oil}) = f(\text{vec(olive)}, \text{vec(oil)})$
  - 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
Previous MWUs Representations

- **Compositional Distributional Representations:**
  - $\text{vec}(\text{olive oil}) = f(\text{vec}(\text{olive}), \text{vec}(\text{oil}))$
  - 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:**
  - \( \text{vec}(\text{olive oil}) = f(\text{vec}(\text{olive}), \text{vec}(\text{oil})) \)
  - 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:**
  - $\mathbf{vec(olive \text{ oil})} = f(\mathbf{vec(olive)}, \mathbf{vec(oil)})$
  - 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
  - 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]

SkipThoughts(s)  What is s’s length?
InferSent(s)  Is w in s?
...

Vered Shwartz and Ido Dagan  •  How well do Pre-trained Text Representations Address Multi-word Units?
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

<table>
<thead>
<tr>
<th>Word Embeddings</th>
<th>Sentence Embeddings</th>
<th>Contextualized Word Embeddings</th>
</tr>
</thead>
<tbody>
<tr>
<td>word2vec</td>
<td>SkipThoughts</td>
<td>ELMo</td>
</tr>
<tr>
<td>GloVe</td>
<td>InferSent*</td>
<td>OpenAI Transformer</td>
</tr>
<tr>
<td>fastText</td>
<td>GenSen*</td>
<td>BERT</td>
</tr>
</tbody>
</table>

* 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:**

```
Authorities meted out summary justice in cases as this
```

```
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

Vered Shwartz and Ido Dagan  •  How well do Pre-trained Text Representations Address Multi-word Units?

9 / 34
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:**
  
  The crash course in litigation made me a better lawyer
2. Noun Compound Literality

Results

Accuracy

Word Embeddings  Sentence Embeddings  Contextualized

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majority</td>
<td>20</td>
</tr>
<tr>
<td>word2vec</td>
<td>26.5</td>
</tr>
<tr>
<td>GloVe</td>
<td>28.8</td>
</tr>
<tr>
<td>fastText</td>
<td>30.3</td>
</tr>
<tr>
<td>SkipThoughts</td>
<td>34.2</td>
</tr>
<tr>
<td>InferSent</td>
<td>24.9</td>
</tr>
<tr>
<td>GenSen</td>
<td>35.5</td>
</tr>
<tr>
<td>ELMo</td>
<td>41.8</td>
</tr>
<tr>
<td>OpenAI Transformer</td>
<td>50</td>
</tr>
<tr>
<td>BERT</td>
<td>44</td>
</tr>
<tr>
<td>Human</td>
<td>87</td>
</tr>
</tbody>
</table>

(1) word embeddings < sentence embeddings < contextualized; (2) Far from humans
2. Noun Compound Literality Analysis

<table>
<thead>
<tr>
<th>ELMo</th>
<th>OpenAI Transformer</th>
<th>BERT</th>
</tr>
</thead>
<tbody>
<tr>
<td>landfill</td>
<td>body</td>
<td>archaeological</td>
</tr>
<tr>
<td>wreckage</td>
<td>place</td>
<td>burial</td>
</tr>
<tr>
<td>Web</td>
<td>man</td>
<td>wreck</td>
</tr>
<tr>
<td>crash</td>
<td>missing</td>
<td>excavation</td>
</tr>
<tr>
<td>burial</td>
<td>location</td>
<td>grave</td>
</tr>
</tbody>
</table>

A search team located the [crash] site and found small amounts of human remains.

After a [crash] 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

The gold/[silver] \( L \) price ratio is often analyzed by traders, investors, and buyers.

<table>
<thead>
<tr>
<th>ELMo</th>
<th>OpenAI Transformer</th>
<th>BERT</th>
</tr>
</thead>
<tbody>
<tr>
<td>silver</td>
<td>platinum</td>
<td>silver</td>
</tr>
<tr>
<td>blue</td>
<td>black</td>
<td>copper</td>
</tr>
<tr>
<td>platinum</td>
<td>gold</td>
<td>platinum</td>
</tr>
<tr>
<td>purple</td>
<td>silver</td>
<td>gold</td>
</tr>
<tr>
<td>yellow</td>
<td>red</td>
<td>diamond</td>
</tr>
</tbody>
</table>

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-literall!
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

![Image](quickmeme.com)
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

Accuracy

Word Embeddings: 50, 60.9, 60.1, 60.7, 51.3, 58.5, 65.6, 67, 50, 74.2, 92
Sentence Embeddings: 60.9, 60.1, 60.7, 51.3, 58.5, 65.6, 67, 50, 74.2, 92
Contextualized

(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

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majority</td>
<td>46.3</td>
</tr>
<tr>
<td>word2vec</td>
<td>41.2</td>
</tr>
<tr>
<td>GloVe</td>
<td>36</td>
</tr>
<tr>
<td>fastText</td>
<td>45.6</td>
</tr>
<tr>
<td>SkipThoughts</td>
<td>47.8</td>
</tr>
<tr>
<td>InferSent</td>
<td>51.5</td>
</tr>
<tr>
<td>GenSen</td>
<td>49.3</td>
</tr>
<tr>
<td>ELMo</td>
<td>43.4</td>
</tr>
<tr>
<td>OpenAI Transformer</td>
<td>52.9</td>
</tr>
<tr>
<td>BERT</td>
<td>50</td>
</tr>
<tr>
<td>Human</td>
<td>77</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Model</th>
<th>F1 Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majority</td>
<td>0</td>
</tr>
<tr>
<td>word2vec</td>
<td>36.6</td>
</tr>
<tr>
<td>GloVe</td>
<td>20.6</td>
</tr>
<tr>
<td>fastText</td>
<td>40.4</td>
</tr>
<tr>
<td>SkipThoughts</td>
<td>23.4</td>
</tr>
<tr>
<td>InferSent</td>
<td>48.4</td>
</tr>
<tr>
<td>GenSen</td>
<td>55.2</td>
</tr>
<tr>
<td>ELMo</td>
<td>45.2</td>
</tr>
<tr>
<td>OALT</td>
<td>14.7</td>
</tr>
<tr>
<td>BERT</td>
<td>37.2</td>
</tr>
<tr>
<td>Human</td>
<td>74.4</td>
</tr>
</tbody>
</table>

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:**

- **VPC**
  - We did get on together
  - Which response did you get on that?

- **Non-VPC**
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’

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


