WHEN POLYSEMY IS WHAT A CONSTRUCTION IS (ALL) ABOUT: Exploring the use of BERT for semantic search and classification

Lauren Fonteyn
Leiden University Centre for Linguistics (LUCL)
Even in historical linguistics, the amount of electronically searchable data has increased substantially over the years.

Growing corpus sizes have had positive consequences for data analysis, but also indirectly affected our modus operandi with regard to the precision-recall trade-off:

“With the increasing availability of very large corpora (…), corpus linguists handle datasets whose size is a continuous challenge to human inspection. (…) Corpus linguists tend to maximize recall (…), which undermines precision. When precision is not optimal, the linguist has to filter the output manually. The larger the corpus and the broader the query, the larger the dataset and the more tedious the clean up. If the dataset is too large, its manual annotation becomes infeasible.”

(Desagulier 2019)
A case study: [BE into X] <> ‘fond of’

(1) Okay, so not every guy is into football. Some love basketball, baseball, or even luge … (COHA, 1998)

(2) A little less conversation and a little more touch my body, cause I’m so into you … (Ariana Grande, 2016)

(3) I tend to like the stuff the rock groups are doing because they're creative and original, and that's something I'm very much into. (OED, 1969)

(4) If the youth public that is so into peace and beatitude were not titillated out of its tepees by this specter of Sodom … (COHA, 1959)
### [BE into X] < > all senses

<table>
<thead>
<tr>
<th>Spatial</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>a) <strong>basic:</strong></td>
<td>- But you’ve been <em>into</em> his room?</td>
</tr>
<tr>
<td>b) <strong>metonymy:</strong></td>
<td>- One of the rebels was <em>into</em> his backpack now</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Metaphorical</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>a) <strong>time:</strong></td>
<td>- I was <em>into</em> my second semester of college</td>
</tr>
<tr>
<td>b) <strong>debt:</strong></td>
<td>- I’m <em>into</em> the boys for quite a few bucks</td>
</tr>
<tr>
<td>c) <strong>sit/act:</strong></td>
<td></td>
</tr>
<tr>
<td>i. <strong>state:</strong></td>
<td>- Some of them are <em>into</em> deep shit</td>
</tr>
</tbody>
</table>
| ii. **activity:** | - Twelve men was *into* that scrap [i.e. ‘fight’]  
- He’s *into* his final medley, with a dozen powerful amps screaming  
- I was deep *into* a John le Carré novel |
[BE into X] < > all senses

2. Metaphorical (continued)
   
d) involvement:
   
i. general: - So Czesko figures this guy must be into some kinda black magic
   
   ii. belief: - I didn’t know you were into conspiracy theories
   
   iii. addiction: - She was really into drugs. She’d lost her apartment and was living in this old van.

   e) liking/fondness:
   
i. general: - My dad is really into watching the travel channel.
   - Keesha and Toya were into hip-hop.
   
   ii. romantic/sexual: - Melanie’s into some wild stuff.
   - It was so refreshing to know that he was into me
I'm all about the blindfold. There's something intensely sensual about not knowing where you're going. (COHA, 2005)

Because you know I'm all about that bass (Meghan Trainor, 2014)

Ya see I'm all about makin that cold cold cash (OED, 1980)

Camille remained unruffled. "Isn't that what we are about, painting our impressions? How else do we express ourselves? (COHA, 1985)
<table>
<thead>
<tr>
<th></th>
<th>Definition</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Spatial</td>
<td>1. He did not appear to be about the house anywhere.</td>
</tr>
<tr>
<td>2.</td>
<td>Approximation</td>
<td>2. I was about thirteen when I started letting the boys feel me up.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- He complained of a headache, that’s about it.</td>
</tr>
<tr>
<td>3.</td>
<td>(Approx.) future</td>
<td>3. I was just about to make some popcorn</td>
</tr>
<tr>
<td>4.</td>
<td>Metaphorical</td>
<td>4. Our discussion was about home-runs</td>
</tr>
<tr>
<td></td>
<td>a) topic</td>
<td>- Hiking is about maintaining a steady pace</td>
</tr>
<tr>
<td></td>
<td>b) concerned with</td>
<td>- Shopping isn’t just about buying a dress</td>
</tr>
<tr>
<td></td>
<td>c) point/purpose</td>
<td>- My mother’s priority was family. She was about family.</td>
</tr>
<tr>
<td></td>
<td>d) like</td>
<td>- I’m all about being friendly and open</td>
</tr>
</tbody>
</table>
[BE PREP X] <-> ‘fond of’

1. **BE INTO**
   - Hits in COHA for [BE (ADV)(ADV)(ADV) into]: 1,683 (1820 - 2010)
   - Relevant hits: 874 ~52%

2. **BE ABOUT**
   - Hits in COHA for [BE (ADV)(ADV)(ADV) into]: 55,273 (1940 - 2010)
   - Relevant hits: 7,375 ~13%
Meanwhile, the increasing size of language corpora allowed computational linguists to firmly establish that semantically similar words tend to have similar contextual distributions (Smith 2019; in reference to e.g. Firth 1957; Deerwester et al. 1990; Miller and Charles 1991).

This led to the development of DSMs, which use vectors (strings of real numbers) that keep track of the contexts in which words appear in a large corpus as proxies for meaning representations (Clark, 2013; Erk, 2012; Turney and Pantel, 2010).
Meanwhile, the increasing size of language corpora allowed computational linguists to firmly establish that semantically similar words tend to have similar contextual distributions (Smith 2019; in reference to e.g. Firth 1957; Deerwester et al. 1990; Miller and Charles 1991).

This led to the development of DSMs, which use vectors (strings of real numbers) that keep track of the contexts in which words appear in a large corpus as proxies for meaning representations (Clark, 2013; Erk, 2012; Turney and Pantel, 2010).

**Distributional Semantic Models (DSM)**

- **cat**: 0.6 0.9 0.1 0.4 -0.7 -0.3 -0.2
- **kitten**: 0.5 0.8 -0.1 0.2 -0.6 -0.5 -0.1
- **dog**: 0.7 -0.1 0.4 0.3 -0.4 -0.1 -0.3
- **houses**: -0.8 -0.4 -0.5 0.1 -0.9 0.3 0.8

Dimensionality reduction of word embeddings from 7D to 2D.
Initially, word vectors allowed us to capture the (changing) meaning of word types (Count models: e.g. Jentet 2013, Perek 2016; Context-predicting: word2vec, e.g. Hamilton et al. 2016, Rosin et al. 2017)

However:

“Typically, a vector matrix assigns one vector representation per word. All the meanings derived from the word distribution in the corpus are conflated into a single string of real numbers. In [the case of homonymy], conflating unrelated meanings into a single vector makes little sense.” (Desagulier 2019)

In some cases it also makes little sense to work with type vectors in cases polysemy: the single vector for each word is forced to represent this wide range of meanings, rendering it virtually useless to detect senses of words (cf. Cook et al. 2014).
Contextualized Token Embeddings

The solution: vectors, but for tokens

Token-based semantic vector spaces based on co-occurrence matrixes

- e.g. Sagi et al. (2011), Heylen et al. (2012), Hilpert & Correia Saavedra (2017)

**Note:** “token-based semantic vectors spaces are constructed by matching the words found in concordance lines with frequency vectors from a type-based semantic vector space” (Hilpert & Correia Saavedra 2017: 3)

Context-predicting models yielding contextualized token embeddings

- e.g. ELMo (Peters et al. 2018), BERT (Devlin et al. 2018)

These methods train a complex, deep neural network to map a vector to each word token based on the entire surrounding context.
These context-predicting models (ELMo, ULMFiT, BERT) are **bidirectional**:

- Most bidirectional LSTMs train a standard left-to-right language model and also train a right-to-left (reverse) language model that predicts previous words from subsequent words: e.g. ELMo uses a single LSTM for the forward language model and backward language model each.

- BERT is different as it takes both directions into account at the same time using a masking procedure.
Can context-predicting models such as BERT help us to more efficiently retrieve the relevant uses of homonymous/polysemous constructions? In other words, can we use these models to conduct semantic searches?

Can these models reveal something about how (closely) the different senses of polysemous words/constructions are related?

Is there a correlation between BERT’s uncertainty in distinguishing (sub)senses, and ambiguity as recognized by human annotators?
Linguistic data retrieval and analysis with BERT
1. Get BERT
   - [Pytorch] https://github.com/huggingface/pytorch-transformers
   - [Tensorflow] https://github.com/hanxiao/bert-as-service
2. Model is pre-trained, but if needed the model can be fine-tuned (on COHA).
3. BERT can be used to make embeddings of [your data set], in this case: instances of [BE (ADV) into|about] as retrieved from COHA.
4. These BERT sentence embeddings can be represented as a matrix in which the rows represent the vectors (of $n$ dimensions) of all contextualized words in the sentence. Rather than pooling the rows into a sentence embedding, we are interested in the contextualized token embeddings of into and about. Determine further settings for embeddings (e.g. context, layer, etc.).
Model & Embeddings
- BERT_{BASE} - Fine-tuned: COHA (1950-2010), 10 epochs
- context: 20 - Layer: -2

Data set
- 588 examples of \([BE \ ADV \ into]\]
  - manual annotation for evaluation:
    - 387 metaphorical (all subtypes)
    - 201 other (spatial, irrelevant)
- 1010 examples of \([BE \ ADV \ about]\]
  - manual annotation for evaluation:
    - 243 metaphorical (all subtypes)
    - 767 other (spatial, approx., future, irrelevant)
**TASK 1**
Distinguish metaphorical from other uses

**TASK 2**
Distinguish subtypes within metaphorical uses
**TASK 1**

Distinguish metaphorical from other (e.g. spatial, future) uses

**Evaluation**

mean average precision (meta): 0.97

classifier accuracy (f1): 0.99
**TASK 2**

Distinguish subtypes within metaphorical uses

**Evaluation**

mean average precision (point_purpose_concern): 0.77

classifier accuracy (f1): 0.88

[Image of t-SNE plot with subtypes: concerned_with, topic, point_purpose, like, structure cluster]
• Structure cluster:

and tattoos, then don't take Dan. That's not what he's about.

sentence: " // Dickau's fascination with Stockton runs from basketball cards (as a

label: meta

subtype: like

order: 1010
**TASK 1**

Even a completely unsupervised application of BERT\textsubscript{BASE} embeddings is successful in distinguishing different uses of [BE \textit{about}].

**TASK 2**

Distinguishing subtypes within the metaphorical group is less clear-cut. Two important patterns emerge:

1. BERT finds \textit{concerned_with} and \textit{point_purpose} difficult to distinguish (\textbf{but so do II!})
2. BERT distinguishes a structure cluster of \textit{wh}-clefts.
**TASK 1**

Distinguish non-metaphorical (i.e. spatial) from metaphorical uses

**Evaluation**

mean average precision (meta): 0.91

classifier accuracy (f1): 0.97
**TASK 2**

Distinguish subtypes within metaphorical uses

**Evaluation**

mean average precision (like/involvement): 0.91

classifier accuracy (f1): 0.94
**TASK 2**

Distinguish subtypes within metaphorical uses

**Evaluation**

mean average precision (like/involvement): 0.91

classifier accuracy (f1): 0.94
**TASK 2**

> **time**

*He was into his early eighties*

> **temporal situation/activity**

*By the time she was on her feet she was already into her dance*

> **habitual/atemporal activity**

*You’re deep into cycling when you care about lube.*
TASK 1

Even a completely unsupervised application of BERT_{BASE} embeddings is quite successful in recognizing metaphorical uses of [BE into].

TASK 2

Distinguishing subtypes within the metaphorical group is a bit less clear-cut. Two important patterns emerge:

1. BERT finds involvement and liking/fondness difficult to distinguish (again, so do I!)
2. The BERT embeddings seem to reveal a gradient from time > temporal situation/activity > habitual/atemporal activity
Conclusions
ABOUT - PCA

topic > concerned with / point_purpose > like

INTO - PCA

time > act_sit > involvement / like
Conclusions

Can context-predicting models such as BERT help us to more efficiently retrieve the relevant uses of homonymous/polysemous constructions? In other words, can we use these models to conduct semantic searches?

Yes.

Can these models reveal something about how (closely) the different senses of polysemous words/constructions are related?

Yes.

Is there a correlation between BERT’s uncertainty in distinguishing (sub)senses, and ambiguity as recognized by human annotators?

Yes.
Conclusions

Limitations:
- BERT is pre-trained on Present-day English data;
- Training BERT on your own data (e.g. Middle English) requires amounts of data (and money / computational power) us mere mortals simply do not have;
- A more realistic option is to fine-tune BERT, for instance on COHA (though you will probably need a ‘plus-sized’ GPU).

However, things are developing fast: similar -- but lighter / cheaper / more easily ‘trainable’ -- models with comparable performance are a-coming.
THANK YOU.

HELLO PLEASE COME TO ICEHL-21 IN LEIDEN
SUBMISSION DEADLINE EXTENDED TO SEPTEMBER 10
SEE OUR TWITTER: @ICEHL21
REFERENCES

REFERENCES