What do neural models tell us about the nature of language?

Juan Luis Gastaldi
ETH Zürich (D-GESS, Turing Centre)
juan.luis.gastaldi@gess.ethz.ch
There is no “philosophy” of language. There is only linguistics.

Louis Hjelmslev

Principes de Grammaire Générale, 1928
DNN AND NATURAL LANGUAGE

\[ \text{i-th output} = P(w_t = i \mid \text{context}) \]

Source: Bengio et al., 2003
Juan Luis Gastaldi | What do neural models tell us about the nature of language?

Mikolov et al., 2013
**Word Embeddings**

* a cat *catches* a mouse

Source: Ferrone et al., 2017
Juan Luis Gastaldi | What do neural models tell us about the nature of language?

Mikolov et al., 2013.

https://projector.tensorflow.org

Hamilton et al., 2016.

Hewitt et al., 2019.
\[ \mathbf{v}_{\text{house}} - \mathbf{v}_{\text{city}} + \mathbf{v}_{\text{countryside}} \approx \mathbf{v}_{\text{farmhouse}} \]

\[ \mathbf{v}_{\text{king}} - \mathbf{v}_{\text{man}} + \mathbf{v}_{\text{woman}} \approx \mathbf{v}_{\text{queen}} \]

\[ \mathbf{v}_{\text{king}} - \mathbf{v}_{\text{queen}} \approx \mathbf{v}_{\text{man}} - \mathbf{v}_{\text{woman}} \]
## Word Embeddings

<table>
<thead>
<tr>
<th>Type of relationship</th>
<th>Word Pair 1</th>
<th>Word Pair 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Common capital city</td>
<td>Athens, Greece</td>
<td>Oslo, Norway</td>
</tr>
<tr>
<td>All capital cities</td>
<td>Astana, Kazakhstan</td>
<td>Harare, Zimbabwe</td>
</tr>
<tr>
<td>Currency</td>
<td>Angola, Chicago</td>
<td>Iran, California</td>
</tr>
<tr>
<td>City-in-state</td>
<td>brother, Illinois</td>
<td>Stockton, rial</td>
</tr>
<tr>
<td>Man-Woman</td>
<td>sister, grandson</td>
<td>Zuliana, granddaughter</td>
</tr>
<tr>
<td>Adjective to adverb</td>
<td>apparent, rapidly</td>
<td>Ethical, unethical</td>
</tr>
<tr>
<td>Opposite</td>
<td>possibly, greater</td>
<td>Tough, tougher</td>
</tr>
<tr>
<td>Comparative</td>
<td>great, easier</td>
<td>Lucky, luckiest</td>
</tr>
<tr>
<td>Superlative</td>
<td>easy, thinking</td>
<td>Read, reading</td>
</tr>
<tr>
<td>Present Participle</td>
<td>think, thinking</td>
<td>Cambodia, Cambodian</td>
</tr>
<tr>
<td>Nationality adjective</td>
<td>Switzerland, Swiss</td>
<td>Walking, swimming</td>
</tr>
<tr>
<td>Past tense</td>
<td>walking, walked</td>
<td>Dollar, dollars</td>
</tr>
<tr>
<td>Plural nouns</td>
<td>mouse, mice</td>
<td>Speak, speaks</td>
</tr>
<tr>
<td>Plural verbs</td>
<td>work, works</td>
<td></td>
</tr>
</tbody>
</table>

Source: Mikolov et al., 2013.
1. The automatic reconstruction of the underlying organization of language does not require more human intervention than the one implied in the most ordinary use of language as recorded in a practically raw linguistic corpus.

2. In that reconstruction, both semantic and syntactic contents of words are determined at once and as the result of the same procedure.

3. Word vector representations are not simply disposed in similarity neighbourhoods, but that the vector space itself is also structured following precise directions at the crossroads of which syntactic and semantic contents are established.
Recent Evolution of Neural NLP Models

- Word Embeddings (word2vec, GloVe)
- Recurrent Neural Nets (LSTM)
- Encoding-Decoding (seq2seq)
- Transformers (attention, GPT-3)
Juan Luis Gastaldi | What do neural models tell us about the nature of language?

Christopher Potts (@ChrisGPotts):
Introducing ML into semantics leads to a reassessment of distinctions like sentence/speaker meaning

Jacob Andreas (@jacobandreas):
Are sentence embeddings "meaning representations"?
Most of the work I've seen is more about syntactic phenomena

Emily M. Bender (@emilymbender):
What's the task the encodings are learned from?
If it's language modeling, that's *not* a representation of meaning.
The Distributional Hypothesis

• “You shall know a word by the company it keeps!” (Firth, 1957)
• “Words which are similar in meaning occur in similar contexts” (Rubenstein & Goodenough 1965)

• “Words with similar meanings will occur with similar neighbors if enough text material is available” (Schütze & Pedersen 1995)
• “A representation that captures much of how words are used in natural context will capture much of what we mean by meaning” (Landauer & Dumais 1997)

• “Words that occur in the same contexts tend to have similar meanings” (Pantel 2005)
• “The degree of semantic similarity between two linguistic expressions A and B is a function of the similarity of the linguistic contexts in which A and B can appear” (Lenci, 2010)
• Theory of (linguistic) meaning as “usage” (Wittgenstein)
  – “the meaning of a word is defined by the circumstances of its use” (Manning & Schütze, 1999)

• Two versions of the DH:
  – Weak: Correlation between context and word meaning (Spence & Owens, 1990)
  – Strong: Causality attributed to contextual distributions (Miller & Charles, 1991)

• Context: the domain or scope within which entities of the same nature can be presented together (“co-occur”), in such a way that they can be associated by a cognitive agent.
Word Embeddings as Matrix Factorization

Source: topicmodels.west.uni-koblenz.de
**Word Embeddings as Matrix Factorization**

<table>
<thead>
<tr>
<th></th>
<th>…</th>
<th>w</th>
<th>x</th>
<th>y</th>
<th>z</th>
<th>…</th>
</tr>
</thead>
<tbody>
<tr>
<td>…</td>
<td>…</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>…</td>
</tr>
<tr>
<td>a</td>
<td>…</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>…</td>
</tr>
<tr>
<td>b</td>
<td>…</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>…</td>
</tr>
<tr>
<td>c</td>
<td>…</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>…</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>…</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Context</th>
<th>Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>(w, x)</td>
<td>(a, c)</td>
</tr>
<tr>
<td>(w, y)</td>
<td>(a, b, c)</td>
</tr>
<tr>
<td>(w, z)</td>
<td>(b, c)</td>
</tr>
<tr>
<td>(x, y)</td>
<td>(a, b)</td>
</tr>
<tr>
<td>(y, z)</td>
<td>(a, b, c)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Term</th>
<th>Context</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a, c)</td>
<td>(w, x, y, z)</td>
</tr>
<tr>
<td>(a, b, c)</td>
<td>(w, x, y, z)</td>
</tr>
<tr>
<td>(b, c)</td>
<td>(w, y, z)</td>
</tr>
<tr>
<td>(a, b)</td>
<td>(x, y, z)</td>
</tr>
<tr>
<td>(a, b, c)</td>
<td>(w, x, y, z)</td>
</tr>
</tbody>
</table>
**Word Embeddings as Matrix Factorization**

<table>
<thead>
<tr>
<th>a = your c = my</th>
<th>w = apartment  x = house  y = chair  z = stool</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th></th>
<th>w</th>
<th>x</th>
<th>y</th>
<th>z</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>b</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>c</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

*your : house*

*my : apartment*
“The day she came to your house in Paris”

“She bought a house and a bungalow”

“Vous vous trompez”
• Saussure, Hjelmslev, Harris…

(Saussure, 1916)

(Hjelmslev, 1935)

(Saussure, 1916)

(Harris, 1951)
## Structuralism

Juan Luis Gastaldi | What do neural models tell us about the nature of language?

(Harris, 1951)
- F. de Saussure
- R. Jakobson
- L. Hjelmslev
- L. Bloomfield
- Z. Harris

• Linguistic content (including essential aspects of linguistic meaning) is the effect of a virtual structure of classes and dependencies at multiple levels, underlying (and derivable from) the mass of things said or written in a given language.

• The task of linguistic analysis is not just that of identifying loose similarities between words out of distributional properties of a corpus, but rather this other one of explicitly drawing from that corpus the system of strict dependencies between implicit linguistic categories.
A priori it would seem to be a generally valid thesis that for every process there is a corresponding system, by which the process can be analyzed and described by means of a limited number of premises. It must be assumed that any process, can be analyzed into a limited number of elements recurring in various combinations. Then, on the basis of this analysis, it should be possible to order these elements into classes according to their possibilities of combination. And it should be further possible to set up a general and exhaustive calculus of the possible combinations.

Hjelmslev (1943)
Paradigm Derivation

1 2 3 4

... the boy came home ...
one girl went down
no man ran in

: : : :
Juan Luis Gastaldi | What do neural models tell us about the nature of language?

Paradigm Derivation

```
1  2  3  4
...
the boy came home ...
one girl went down
no man ran in
sky

sadness

the have
from
```

- Colorless green ideas sleep furiously ✓
- Furiously sleep ideas green colorless ✗
<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>the</td>
<td>boy</td>
<td>came</td>
</tr>
<tr>
<td>one</td>
<td>girl</td>
<td>went</td>
<td>home</td>
</tr>
<tr>
<td>no</td>
<td>man</td>
<td>ran</td>
<td>down</td>
</tr>
<tr>
<td>sky</td>
<td></td>
<td>in</td>
<td></td>
</tr>
<tr>
<td></td>
<td>sadness</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>gavagai</td>
<td></td>
</tr>
</tbody>
</table>
Juan Luis Gastaldi | What do neural models tell us about the nature of language?
Paradigm Derivation

... the one boy girl came home down...

1 2 3 4

- 0.09000 -
- 0.08000 -
- 0.07000 -
- 0.06000 -
- 0.05000 -
- 0.04000 -
- 0.03000 -
- 0.02000 -
- 0.01000 -
- 0.00000 -

1800 1820 1840 1860 1880 1900 1920 1940 1960 1980 2000

man

Juan Luis Gastaldi | What do neural models tell us about the nature of language? 31
Distillation

Unlabeled samples → Teacher Model → Soft Targets → Distillation Loss

Student Model → Soft Targets → MSE

PARADIGM DERIVATION

... the boy came home ...  
one girl went down 
no man ran in
PARADIGM DERIVATION

... the boy came home
one girl went down
no man ran in...

Juan Luis Gastaldi | What do neural models tell us about the nature of language? 34
PARADIGM DERIVATION

A  B  C  D

... the boy came home ...
one girl went down
no man ran in
the girl ran home
### Non-Neural Distributional Model

<table>
<thead>
<tr>
<th>Procedure Step</th>
<th>Neural</th>
<th>Non-Neural</th>
</tr>
</thead>
<tbody>
<tr>
<td>Segmentation</td>
<td>Sub-word Tokenization, Character level DNNs</td>
<td>BPE</td>
</tr>
<tr>
<td>Classification</td>
<td>Word Embedding Distillation</td>
<td>Matrix models</td>
</tr>
<tr>
<td>Dependencies</td>
<td>Attention</td>
<td>Biorthogonal Typing, Linear Logic</td>
</tr>
</tbody>
</table>
Some Conclusions

• Generic model
• Non-necessarily neural
• Producing explicit representations
• Supporting logical relations
• Capturing significant aspects of meaning
[Link](#)


• Mikolov, Tomáš, Quoc V. Le, and Ilya Sutskever (2013). “Exploiting Similarities among Languages for Machine Translation”. In: CoRR abs/1309.4168.


What do neural models tell us about the nature of language?

Juan Luis Gastaldi
ETH Zürich (D-GESS, Turing Centre)
juan.luis.gastaldi@gess.ethz.ch