Semantic Models of Competence and Performance: either or both?

Raffaella Bernardi
University of Trento

Workshop on
Formal and Distributional Perspectives on Meaning
## Competence vs. Performance

<table>
<thead>
<tr>
<th>Ingredients</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>250g butter, softened</td>
<td>1. Mix 250g softened butter and 140g caster sugar in a large bowl with a wooden spoon, then add 1 egg yolk and 2 tsp vanilla extract and briefly beat to combine. Sift over 300g plain flour and stir until the mixture is well combined – you might need to get your hands in at the end to give everything a really good mix and press the dough together.</td>
</tr>
<tr>
<td>140g caster sugar</td>
<td></td>
</tr>
<tr>
<td>1 egg yolk</td>
<td></td>
</tr>
<tr>
<td>2 tsp vanilla extract</td>
<td></td>
</tr>
<tr>
<td>300g plain flour</td>
<td></td>
</tr>
</tbody>
</table>
“Most formal semanticists who are linguists are very much concerned with human semantic competence. [..]

What is semantic competence? For formal semanticists, [..] given a sentence in a context, and given idealized omniscience [..] semantic competence is widely considered to consist in knowledge of truth conditions and entailment relations of sentences of the language.”
Distributional Semantics (DS): Performance not Competence

Landauer and Dumais 1997

Model human learning process:
• Learning word meaning from data (co-occurrences)
• Generalize evidence (weighting)
• Induce new knowledge (dimensionality reduction)

Evaluate models against human performance on some tasks:
• TOEFL test
Why I have “moved” to Distributional Semantics

Why I have started?
• Because I met Massimo Poesio and Marco Baroni who were working on it.
• Because I couldn’t understand it, hence I got curious.

Why I have continued for so many years?
• Because there is something in it I like a lot and was not there in my studies of FS.
DS main ingredients

Continuous representations (vectors)

Building blocks:
• Semantic space
• Representations learned from lots of data.
• Similarity measure

Tasks:
• Lexical relation, categorization, priming etc.

Methods
• Tasks on rather big real-life test sets
• Statistically based evaluation measures
FS main ingredients

Symbolic representations
Building blocks:
• The meaning of a sentence is the truth value
• Referential meaning (entities as building blocks)
• Semantic compositionality lead by syntax
• Function application (and abstraction)
Task:
• Reasoning (validity vs. satisfiability) driven by grammatical words.
Methods:
• Clean data (fragments)
• Clean results
Which Semantic Model I like most?

- The one that does not exist yet
- The one that will mix features of both FS and DS models
What I like most of FS: Truth Value

The meaning of *Snow is white* is T/F

✓ I want to keep it.
What I like most of FS: Concepts vs. Entities

Concept/Property:
{m, r, d, ..}

Entity/constant:
m

✓ I want to keep it.
What I like most of FS:
Meaning composition driven by syntax

Ding and Melloni 2015: yes

✓ I want to keep it
What I like most of DS Models

- Focus on a data-driven approach
- Interest in cognitive plausibility
- Experiment/evaluation based on behavioral studies
What I have tried to import into DS from FS

Symbolic representations building blocks:
• The meaning of a sentence is the *truth value*
• Referential meaning (*entities* as building blocks)
• Semantic compositionality lead by syntax
• Function application

Task:
• *Reasoning driven by grammatical words.*

Methods:
• Clean data (fragments)
• Clean results
Evaluation based on behavioral studies: composition

Kintsch (2001):  
*The horse run – gallop*  
*The color run – dissolve*

Baroni and Zamparelli (2010)  
Baroni, Bernardi and Zamparelli, *Frege in Space* In LILT 2014

Lesson Learned: additive models go better than expected – but I still don’t know why.
Evaluation based on behavioral studies: entailment

2014 SICK (Sentence involving Compositional Knowledge).

Given A and B: entail, contradict or neutral?
A: *Two teams are competing in a football match*
B: *Two groups of people are playing football*

A: *The brown horse is near a red barrel at the rodeo*
B: *The brown horse is far from a red barrel at the rodeo*

Bentivogli et al. LREV 2016

Lesson Learned: DS Models can capture entailment relations between phrases, worse at higher level. Problems with coordination involving quantities, comitative constructions
Evaluation based on behavioral studies: negation

Logical Negation: \[ P = T \]
\[ \neg P = F \]

Conversational Negation: \[ \neg P = \{ \text{alternatives to } P \} \]

DSMs account for CN. Cosine similarity a proxy of alternatives:
- \( \text{This is not a dog.. It is a wolf} \)
  \( \text{sim(dog, wolf)} = 0.80 \)
- \( \text{This is not a dog.. It is a screwdriver} \)
  \( \text{sim(dog, screwdriver)} = 0.10 \)

Kruszewski et al In Computational Linguistics 2016

Laura Aina MSc Thesis at ILLC (2017):
Not logical: a distributional semantics account of negated adjectives

Lesson learned: Words have logical and conversational meanings – humans master both.
Evaluation based on behavioral studies: quantifiers

Lexical and Phrase entailment
• ACL 2013: Sim(orchestra,many musicians)
• EACL 2013: all N => some N, some N=/=> all N

Given a sentence, can DSMs learn to predict a quantifier? E.g.
“_____ the electoral votes were for Trump, so he was elected”

On-going work with S. Pezzelle, S. Steinert Threlkeld and J. Szymanik

Lesson Learned: Vectors representations encode some properties of quantifiers that distinguish their uses.
Overall lesson learned on Performance and Competence

Conversational and Logical Meaning:

• From corpora, we obtain the conversational meaning humans use. Don’t expect to get the logical one is not the one we use.

• Yet, if humans are asked to use words’ logical meaning they are able to do so.
What I still miss

FS main ingredients I still miss:
• The meaning of a sentence is the *truth value*
• Referential meaning (*entities* as building blocks)
• Semantic compositionality lead by syntax
• Function application (and abstraction)

Task:
• Reasoning (*validity vs. satisfiability*) driven by grammatical words.

Methods:
• Clean data (fragments)
• Clean results

DS main ingredients I still miss:
*Cognitive plausibility?* What about evidence from neuro-science?
Some recent work on:

**Truth values**

Probabilistic Logic as a bridge between DS and FS models by learning *meaning postulates* probabilities from corpora.

Katrin Erk *In Semantics and Pragmatics* 2016

Sadrzadeh et al.: various work on Compositional DSM based on Frobenius algebra
Some recent work on: reference

A vector representation of proper names:

- Characters of a novel (A. Herbelot, IWCS 2015): re-weighting vectors to produce an individual out of a kind.
- Famous people, locations (G. Boleda et al EACL 2017)
Cognitive Plausibility:
Humans are multimodal

M. Andrew, G. Vigliocco and D. Vinson (2009)
Human semantic representations are derived from an optimal statistical combination of [experiential and language distributions]

Barsalou 2008:
Both from Cognitive Psychology and Cognitive Neuroscience there are evidence that higher cognitive processes (e.g. mapping from concepts to instances, composition of symbols to form complex symbolic expression etc..) engage modal systems. [..] The presence of this multimodal representation makes the symbolic operations possible.
Computer Vision

Classification

Classification + Localization

Object Detection

Instance Segmentation

CAT

CAT

CAT, DOG, DUCK

CAT, DOG, DUCK

Single object

Multiple objects

Again, vectors
Multimodal Models

*Multimodal Distributional Semantics*
Bruni, Tran and Baroni (2014)

<table>
<thead>
<tr>
<th></th>
<th>planet</th>
<th>night</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>moon</td>
<td>10</td>
<td>22</td>
<td>22</td>
<td>0</td>
</tr>
<tr>
<td>sun</td>
<td>14</td>
<td>10</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>dog</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>20</td>
</tr>
</tbody>
</table>

*Combining Language and Vision with a Multimodal Skipgram Model*
Lazaridou, Phan and Baroni (2015)
Multimodal models: Performance on VQA

- What color are her eyes?
- What is the mustache made of?
- How many slices of pizza are there?
- Is this a vegetarian pizza?
My wishes on Truth value, validity vs. satisfiability

Snow is white is T/F

1. I would like to have a model that understand whether a sentence is true or false w.r.t an image
2. I would like to have a model that understand whether two pairs of sentences are in an entailment relation w.r.t a given image.
What we have done: false w.r.t an image

Conclusion: Need of a more fine-grained representation.

Shekar et al. ACL 2017, ICWS 2017
What we are doing: grounding entailment

A boy in a blue uniform is standing next to a boy in a red and a boy in yellow ones and they are holding baseball gloves.

=>

Three boys hold baseball gloves.

A performer plays an instrument for the audience?

The performer has a flute.

With C. Greco, H. Vu, A. Erofeeva and A. Gatt. In progress
My wishes on quantifiers

3. I’d like to have a model that has competence on quantifiers:
   Some girls are eating a pizza

   SOME PIZZA > SOME GIRL

   SOME GIRL > SOME PIZZA

4. I’d like to have a model that use quantifiers as humans do:
   “Hey, someone ate all chocolate”
What we have done: learning quantities from vision

Q. How many pets are cats?
A. Two / Some / 40%

Conclusion: Neural Networks learn to compare sets, assign quantifiers and estimate proportions.

What I would like to study next

Improve the **multimodal representations**, in particular find:

- ways to distinguish in the vector space: *entities vs concepts*  
  (future work with A. Herbelot and G. Boleda)

- ways to *store facts* and *update* multimodal vectors as new knowledge about the entity or concept is gained.  
  (current work with R. Fernandez et al. on Visual Dialogue)

**Go back to Barsalou’s claim:**

“The presence of this multimodal representation makes the **symbolic operations possible.**”

I find the work on the combination of DRT and DSM a possible direction to reach this aim.

McNally and Boleda 2017, ERC AMORE PI: G. Boleda
Back to competence: diagnostic tasks

Figure 1: Example sentences and images from our corpus. Each image includes three boxes with different object types. The truth value of the top sentence is true, while the bottom is false.

Alane Shur, Mike Lewis, James Yeh, and Yoav Artzi. ACL 2017
Evaluation of models on diagnostic datasets

<table>
<thead>
<tr>
<th></th>
<th>VQA (abs)</th>
<th>VQA (real)</th>
<th>Our Data</th>
<th>NMN Correct</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Semantics</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cardinality (hard)</td>
<td>12</td>
<td>11.5</td>
<td>66</td>
<td>63.8</td>
<td>There are exactly four objects not touching any edge</td>
</tr>
<tr>
<td>Cardinality (soft)</td>
<td>0</td>
<td>1</td>
<td>16</td>
<td>63.4</td>
<td>There is a box with at least one square and at least three triangles</td>
</tr>
<tr>
<td>Existential</td>
<td>4.5</td>
<td>11.5</td>
<td>88</td>
<td>64.2</td>
<td>There is a tower with yellow base.</td>
</tr>
<tr>
<td>Universal</td>
<td>1</td>
<td>1</td>
<td>7.5</td>
<td>67.8</td>
<td>There is a black item in every box.</td>
</tr>
<tr>
<td>Coordination</td>
<td>3</td>
<td>5</td>
<td>17</td>
<td>58.5</td>
<td>There are 2 blue circles and 1 blue triangle</td>
</tr>
<tr>
<td>Coreference</td>
<td>8.5</td>
<td>6.5</td>
<td>3</td>
<td>55.3</td>
<td>There is a blue triangle touching the wall with its side.</td>
</tr>
<tr>
<td>Spatial Relations</td>
<td>31</td>
<td>42.5</td>
<td>66</td>
<td>61.6</td>
<td>There is one tower with a yellow block above a yellow block</td>
</tr>
<tr>
<td>Comparative</td>
<td>1.5</td>
<td>1</td>
<td>3</td>
<td>73.6</td>
<td>There is a box with multiple items and only one item has a different color</td>
</tr>
<tr>
<td>Presupposition²</td>
<td>79</td>
<td>80</td>
<td>19.5</td>
<td>54.0</td>
<td>There is a box with seven items and the three black items are the same in shape</td>
</tr>
<tr>
<td>Negation</td>
<td>0</td>
<td>1</td>
<td>9.5</td>
<td>51.0</td>
<td>There is exactly one black triangle not touching the edge</td>
</tr>
<tr>
<td><strong>Syntax</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coordination</td>
<td>0</td>
<td>0</td>
<td>4.5</td>
<td>53.4</td>
<td>There is a box with at least one square and at least three triangles</td>
</tr>
<tr>
<td>PP Attachment</td>
<td>7</td>
<td>3</td>
<td>23</td>
<td>70.9</td>
<td>There is a black block on a black block as the base of a tower with three blocks</td>
</tr>
</tbody>
</table>

Shur et al. ACL 2017

Idealized context allows a clear evaluation of the model.
My wish-list and expectations

I’d like to see a model that

- simulates human competence on very simplified semantic tasks (clean data = diagnostic datasets)
- simulates human performance (both correct and wrong answers) on real life semantic tasks

I’d expect that the model:

- will be multimodal
- will be trained incrementally using Machine Learning
- will combine continuous and symbolic representations
Competence and Performance
Logical and statistical reasoning
People I would like to thank

Aurelie Herbelot, Albert Gatt, Adrian Muscat, Gemma Boleda, Katerina Pastra, Marco Baroni, Manuela Piazza, Massimo Poesio, Raquel Fernandez, Roberto Zamparelli and Sandro Pezzelle

The students of the Computational Linguistics class, University of Trento
CFP: JNLE Special Issue on Representation of Sentence Meaning

O. Bojar, R. Bernardi, H. Schwenk and B. Webber (guest editors)

- Relation between traditional symbolic meaning representations and the learned continuous ones.
- Which properties of meaning representations are most desirable, universally.
- Comparisons of types of meaning representations (e.g. fixed-size vs. variable-length) and methods for learning them.
- Techniques of explorations of learned meaning representations.
- Evaluation methodologies for meaning representations, including surveys thereof.
- Extrinsic evaluation by relations to cognitive processes.
- Broad summaries of psycholinguistic evidence describing properties of meaning representation in the human brain.

Expected submission deadline: October 2018