W2_1: VISUAL GROUNDING AND MULTIMODAL NLP

Cognitive Approaches to Multimodal Language Processing

Sandro Pezzelle
sandropezzelle.github.io

LANGUAGE AND DIALOGUE MODELS: NLU / NLG

Natural Language Understanding

Context: Both its sun-speckled shade and the cool grass beneath were a welcome respite after the stifling kitchen, and I was glad to relax against the tree's rough, brittle bark and begin my breakfast of buttery, toasted bread and fresh fruit. Even the water was tasty, it was so clean and cold. It almost made up for the lack of _____.

Target word: **coffee**

[Paperno et al. 2016; Vinyals & Le 2015]

LANGUAGE AND DIALOGUE MODELS: NLU / NLG

Natural Language Understanding

Context: Both its sun-speckled shade and the cool grass beneath were a welcome respite after the stifling kitchen, and I was glad to relax against the tree's rough, brittle bark and begin my breakfast of buttery, toasted bread and fresh fruit. Even the water was tasty, it was so clean and cold. It almost made up for the lack of _____.

Target word: **coffee**

Natural Language Generation

User: hello?

Bot: hello

User: where am I?

Bot: you're in a hospital.

User: who are you?

Bot: I'm a lawyer.

User: how are you doing?

Bot: I'm fine.

User: are you my friend?

Bot: no.

[Paperno et al. 2016; Vinyals & Le 2015]

LANGUAGE-ONLY SEMANTIC MODELS

Semantics contrasts with syntax, the study of the combinatorics of units of a language (without reference to their meaning), and pragmatics, the study of the relationships between the symbols of a language, their meaning, and the users of the language.[5] Semantics as a field of study also has significant ties to various representational theories of meaning including truth theories of meaning, coherence theories of meaning, and correspondence theories of meaning. Each of these is related to the general philosophical study of reality and the representation of meaning. In 1960s psychosemantic studies became popular after Osgood's massive cross-cultural studies using his semantic differential (SD) method that used thousands of nouns and adjective bipolar scales. A specific form of the SD, Projective Semantics method uses only most common and neutral nouns that correspond to the 7 groups (factors) of adjective-scales most consistently found in cross-cultural studies (Evaluation, Potency, Activity as found by Osgood, and Reality, Organization, Complexity, Limitation as found in other studies). In this method, seven groups of bipolar adjective scales corresponded to seven types of nouns so the method was thought to have the object-scale symmetry (OSS) between the scales and nouns for evaluation using these scales. For example, the nouns corresponding to the listed 7 factors would be: Beauty, Power, Motion, Life, Work, Chaos, Law. Beauty was expected to be assessed unequivocally as “very good” on adjectives of Evaluation-related scales, Life as “very real” on Reality-related scales, etc.

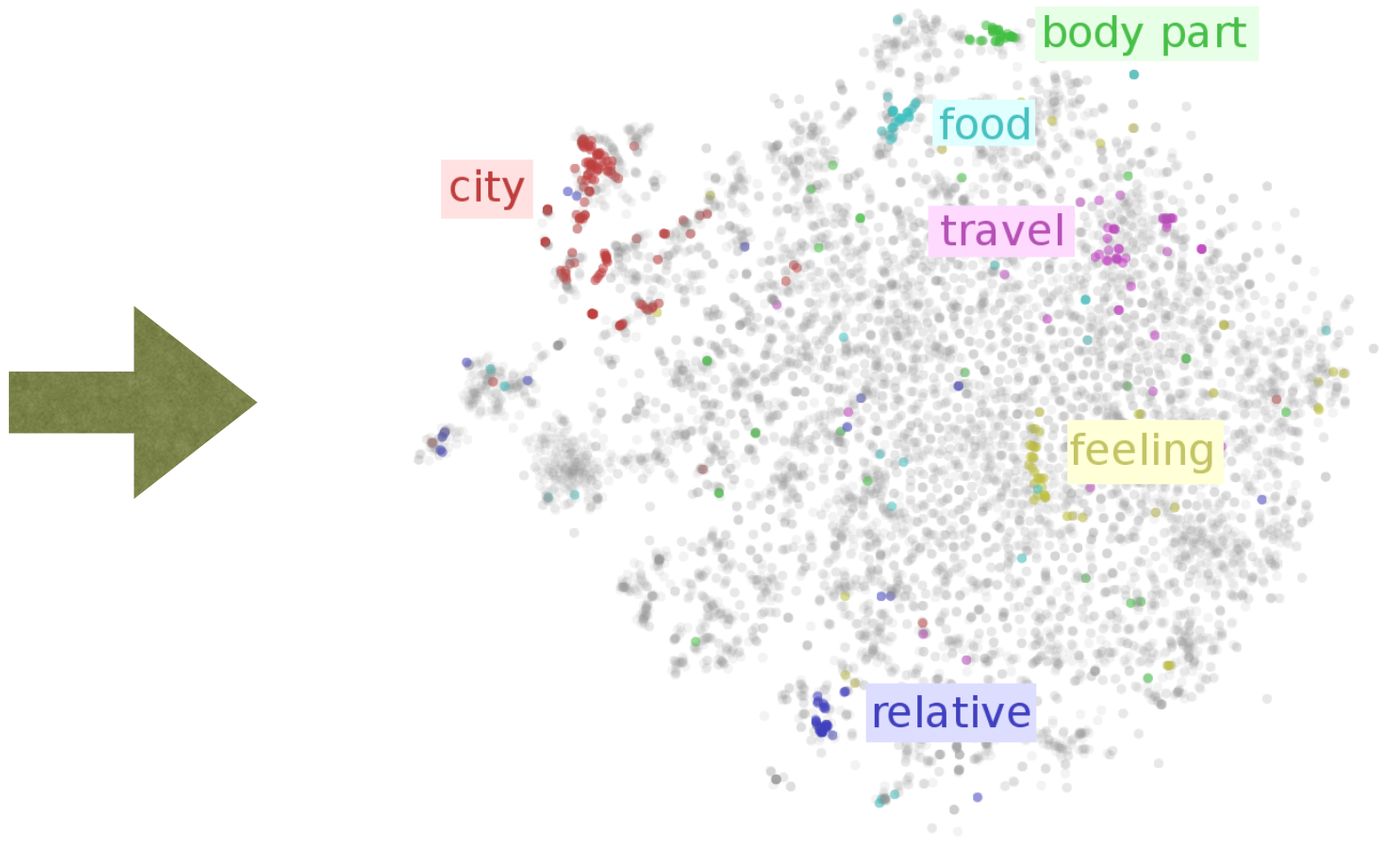


Image credits: https://colah.github.io/

> According to … [language-only semantic models], the *sky* is *green*; *flour* is *black*, and *violins* are *blue*.

–Marco Baroni, 2016

LANGUAGE-ONLY SEMANTIC MODELS: DSMs

- *Distributional hypothesis*: words with **similar meaning** occur in **similar contexts**
- Words as vectors encoding **statistics of text *corpora***
- The closer the vectors in the space, the higher their **similarity**

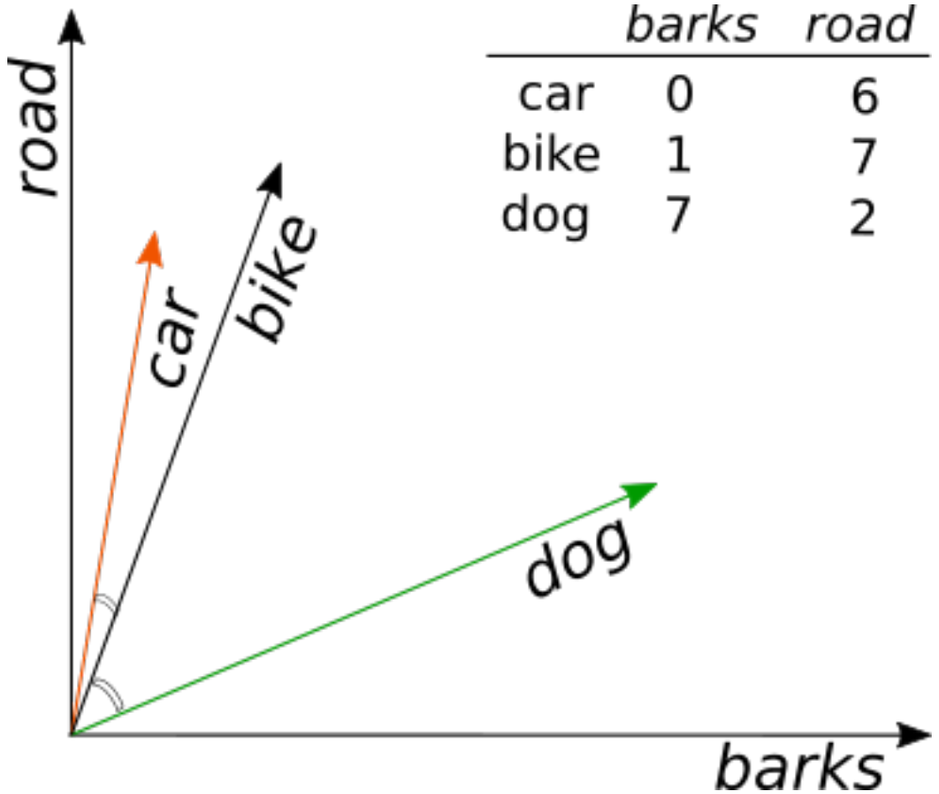
LANGUAGE-ONLY SEMANTIC MODELS: DSMs

raw text

Semantics contrasts with syntax, the study of the combinatorics of units of a language (without reference to their meaning), and pragmatics, the study of the relationships between the symbols of a language, their meaning, and the users of the language.[5] Semantics as a field of study also has significant ties to various representational theories of meaning including truth theories of meaning, coherence theories of meaning, and correspondence theories of meaning. Each of these is related to the general philosophical study of reality and the representation of meaning. In 1960s psychosemantic studies became popular after Osgood's massive cross-cultural studies using his semantic differential (SD) method that used thousands of nouns and adjective bipolar scales. A specific form of the SD, Projective Semantics method uses only most common and neutral nouns that correspond to the 7 groups (factors) of adjective-scales most consistently found in cross-cultural studies (Evaluation, Potency, Activity as found by Osgood, and Reality, Organization, Complexity, Limitation as found in other studies). In this method, seven groups of bipolar adjective scales corresponded to seven types of nouns so the method was thought to have the object-scale symmetry (OSS) between the scales and nouns for evaluation using these scales. For example, the nouns corresponding to the listed 7 factors would be: Beauty, Power, Motion, Life, Work, Chaos, Law. Beauty was expected to be assessed unequivocally as “very good” on adjectives of Evaluation-related scales, Life as “very real” on Reality-related scales, etc.

multi-dimensional semantic space

	barks	*road*
car	0	6
bike	1	7
dog	7	2



LIMITATIONS: THE COLOR EXAMPLE

- DSMs fail to label **objects with their typical color**, e.g., "yellow banana" [Bruni et al. 2012]

- Why? In texts, **no/little mention of prototypical color of things**, e.g., "I ate a yellow banana" or "I play a brown violin"

LANGUAGE-ONLY SEMANTIC MODELS: LIMITATIONS

raw text

Semantics contrasts with syntax, the study of the combinatorics of units of a language (without reference to their meaning), and pragmatics, the study of the relationships between the symbols of a language, their meaning, and the users of the language.[5] Semantics as a field of study also has significant ties to various representational theories of meaning including truth theories of meaning, coherence theories of meaning, and correspondence theories of meaning. Each of these is related to the general philosophical study of reality and the representation of meaning. In 1960s psychosemantic studies became popular after Osgood's massive cross-cultural studies using his semantic differential (SD) method that used thousands of nouns and adjective bipolar scales. A specific form of the SD, Projective Semantics method uses only most common and neutral nouns that correspond to the 7 groups (factors) of adjective-scales most consistently found in cross-cultural studies (Evaluation, Potency, Activity as found by Osgood, and Reality, Organization, Complexity, Limitation as found in other studies). In this method, seven groups of bipolar adjective scales corresponded to seven types of nouns so the method was thought to have the object-scale symmetry (OSS) between the scales and nouns for evaluation using these scales. For example, the nouns corresponding to the listed 7 factors would be: Beauty, Power, Motion, Life, Work, Chaos, Law. Beauty was expected to be assessed unequivocally as "very good" on adjectives of Evaluation-related scales, Life as "very real" on Reality-related scales, etc.

"black"

of uk size 6 adiddas y3 black trainers, they are
of mens Rockport XCS, black soft leather boots
. Happy bidding!! Tina x Black Gardenia 4" inch
new, never before worn black sequin UGG boots
select.not been worn . black flat shoes size 4.
bidding! Never worn. Black Fred perry pumps.
are a pair of barely worn Black Nike Plimsolls. They
great pair of gently worn black leather boots from
(all buckles are working) Black with Gold buckles.
Brand new Women's black and white Converse,
Android app women's black shoe's size 5 Next
pair of women's black Kickers genuine

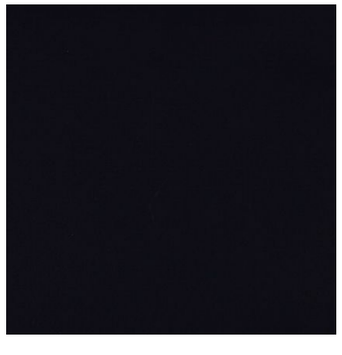
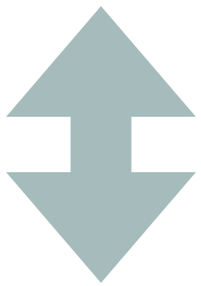


Image credits: https://sites.google.com/site/clexercises/home/concordances

A MISSING LINK

- DSMs fail to label **objects with their typical color**, e.g., "yellow banana" [Bruni et al. 2012]
- Why? In texts, **no/little mention of prototypical color of things**, e.g., "I ate a yellow banana" or "I play a brown violin"
- For people, this information **implicit** thanks to our experience



Humans acquire word meanings in richly situated settings: **linguistic knowledge + external world** perceived through our senses (vision, sound, etc.)

LANGUAGE-ONLY SEMANTIC MODELS: RECENT APPROACHES

- Shallow, context-free (aka "predict" DSMs): `word2vec`

  [Mikolov et al. 2013], `GloVe` [Pennington et al. 2014]

- Deep (biLSMT), contextualized: `ELMo` [Peters et al. 2018]

- Deep (Transformer), contextualized: BERT [Devlin et al. 2019]

SAME LIMITATIONS!

LANGUAGE-ONLY SEMANTIC MODELS: RECENT APPROACHES

- Shallow, context-free (aka "predict" DSMs): `word2vec`

  [Mikolov et al. 2013], `GloVe` [Pennington et al. 2014]

- Deep (biLSMT), contextualized: `ELMo` [Peters et al. 2018]

- Deep (Transformer), contextualized: BERT [Devlin et al. 2019]

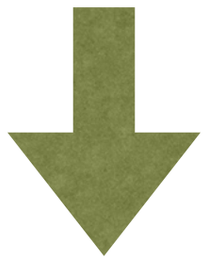
SAME LIMITATIONS!

BUT no consensus/full understanding on whether these representations are **directly comparable** to each other
[Mickus et al. 2019; Westera & Boleda 2019]

THE SYMBOL GROUNDING PROBLEM

- "[These models] follow […] a 'solipsistic' route to semantics in which the meaning of a word is entirely accounted for by **patterns of co-occurrence with other words**" [Baroni 2016]

- No access to the **sensorimotor world** (in contrast, human concepts are strongly *embodied* in our senses [Barsalou 2008])



The symbol grounding problem

[Harnad 1990]

WHY SYMBOL GROUNDING: 2 (MAIN) REASONS

Theoretically/cognitively valid **meaning representations**

- e.g., a representation of “violin” that *includes* that violins are wooden, brown, are played with hands, have a given sound…

Image credits: https://www.musicalhow.com/violin-notes-finger-placement/

WHY SYMBOL GROUNDING: 2 (MAIN) REASONS

Theoretically/cognitively valid **meaning representations**

- e.g., a representation of "violin" that *includes* that violins are wooden, brown, are played with hands, have a given sound…

Usefulness/effectiveness of **real-world applications**

- e.g., the task of learning to execute a natural language instruction in a *situated* context (pick the cup, move right, etc.)

Image credits: https://www.musicalhow.com/violin-notes-finger-placement/

Image credits: https://unsplash.com/s/photos/cup-of-coffee

TODAY: VISUAL GROUNDING

- Symbol grounding ideally involves any **human senses** (visual, auditory, olfactory, haptic, etc.)

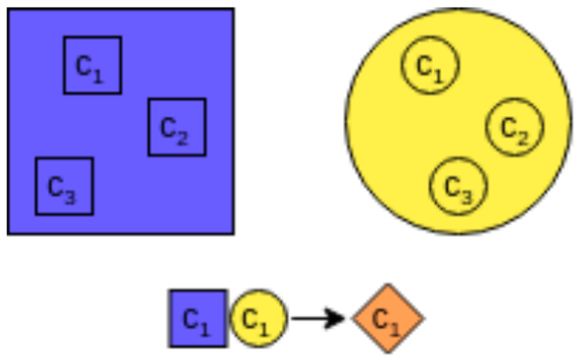
Visual grounding

Link between **language** and what it denotes in the **visual world** ~ an image, a video, etc.

VISUAL GROUNDING: 2 (MAIN) COMPUTATIONAL APPROACHES

Visually-grounded semantics

- **combination** of *known* textual and visual features to obtain **human-like representations** (task-agnostic)



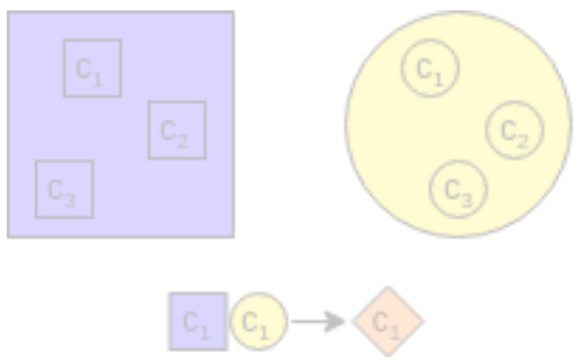
(a) Multimodal fusion. Concatenate known representations from modality A and B and apply dimensionality reduction.

[Beinborn et al. 2018]

VISUAL GROUNDING: 2 (MAIN) COMPUTATIONAL APPROACHES

Visually-grounded semantics

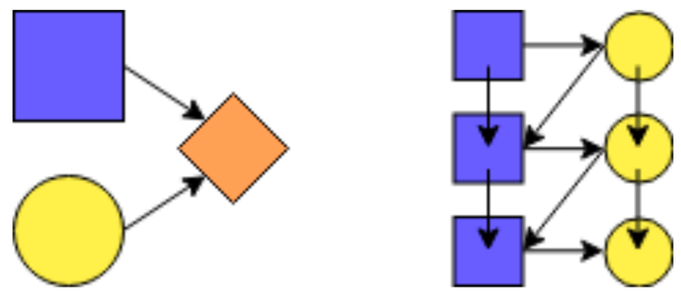
- **combination** of *known* textual and visual features to obtain **human-like representations** (task-agnostic)



(a) Multimodal fusion. Concatenate known representations from modality A and B and apply dimensionality reduction.

Multimodal machine learning (aka Language & Vision tasks)

- **joint multimodal processing** to understand/align/integrate **information from language and vision** (task-oriented)



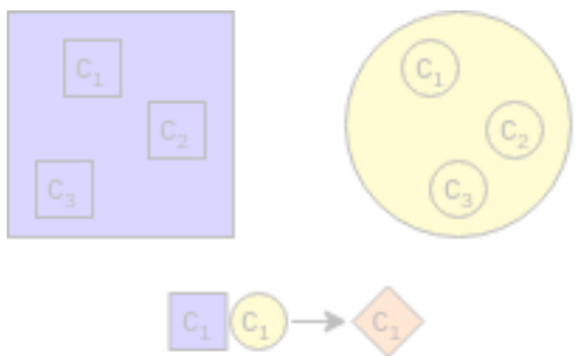
(c) Joint multimodal processing. Left: Modality A and B both contribute to a joint prediction. Right: Interactive exchange of information between modalities.

[Beinborn et al. 2018]

VISUAL GROUNDING: 2 (MAIN) COMPUTATIONAL APPROACHES

Visually-grounded semantics

- **combination** of *known* textual and visual features to obtain **human-like representations** (task-agnostic)

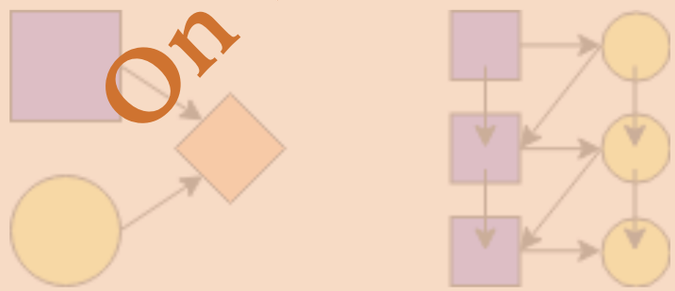


(a) Multimodal fusion. Concatenate known representations from modality A and B and apply dimensionality reduction.

Multimodal machine learning (aka Language & Vision tasks)

- **joint multimodal processing** to understand/align/integrate **information from language and vision** (task-oriented)

On Thursday!



(c) Joint multimodal processing. Left: Modality A and B both contribute to a joint prediction. Right: Interactive exchange of information between modalities.

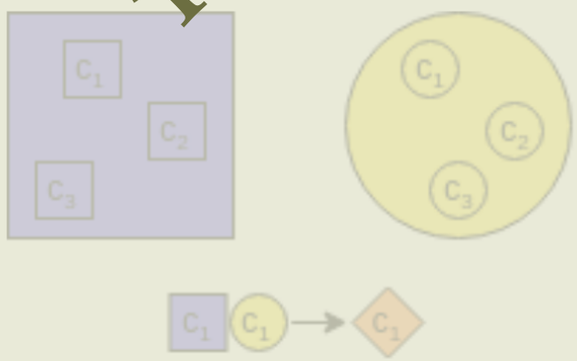
[Beinborn et al. 2018]

VISUAL GROUNDING: 2 (MAIN) COMPUTATIONAL APPROACHES

Visually-grounded semantics

- **combination** of *known* textual and visual features to obtain **human-like representations** (task-agnostic)

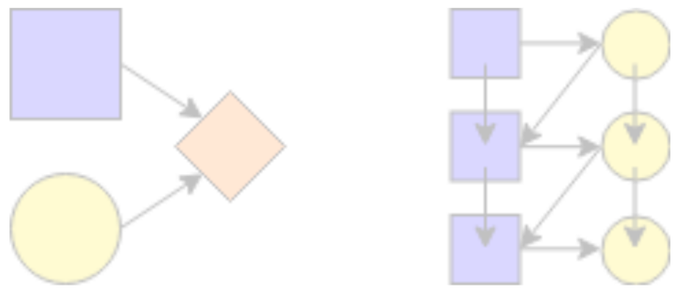
Today!



(a) Multimodal fusion. Concatenate known representations from modality A and B and apply dimensionality reduction.

Multimodal machine learning (aka Language & Vision tasks)

- **joint multimodal processing** to understand/align/integrate **information from language and vision** (task-oriented)



(c) Joint multimodal processing. Left: Modality A and B both contribute to a joint prediction. Right: Interactive exchange of information between modalities.

[Beinborn et al. 2018]

LET'S HAVE A BREAK! SEE YOU AT 3PM