On the Status of Word Embeddings as Implementations of the Distributional Hypothesis

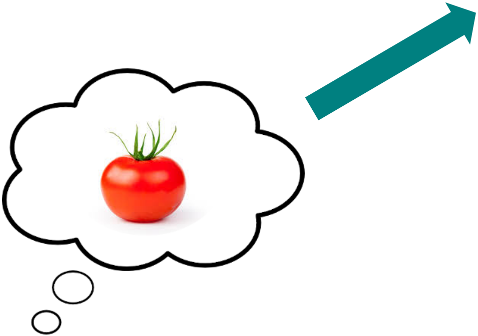
Timothee Mickus

Jury Members

Supervisor	Mathieu Constant, Université de Lorraine
Co-supervisor	Denis Paperno, Universiteit Utrecht
Reviewers	Benoit Crabbé, Université de Paris
	Nabil Hathout, CNRS / Université de Toulouse Jean Jaurès
Examiners	Gemma Boleda, Universitat Pompeu Fabra
	Vera Demberg, Universität des Saarlandes
	Claire Gardent, CNRS / Université de Lorraine
	Alessandro Lenci, Università di Pisa
Guest member	Kees van Deemter, Universiteit Utrecht

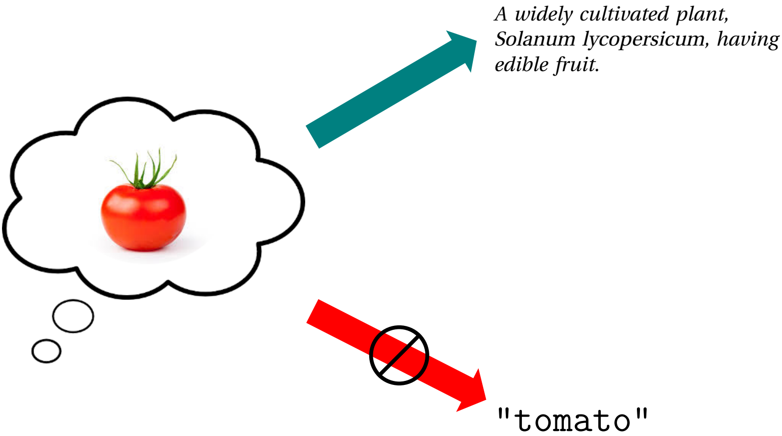
Meaning?

Meaning?

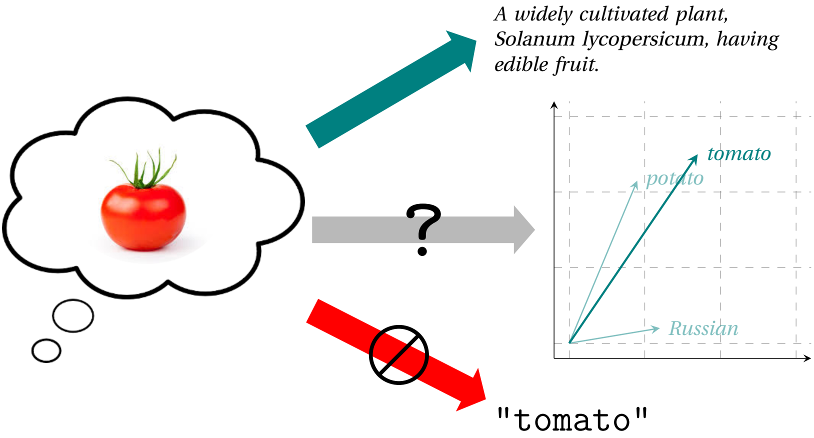


A widely cultivated plant, Solanum lycopersicum, having edible fruit.

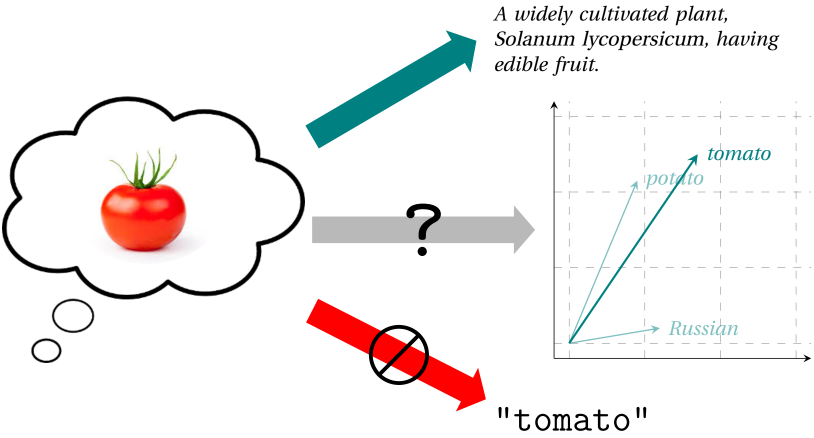
Meaning?


A widely cultivated plant, Solanum lycopersicum, having edible fruit.
"tomato"

Meaning?


A widely cultivated plant, Solanum lycopersicum, having edible fruit.
?
tomato
potato
Russian
"tomato"

Meaning?



Today's talk:

- **Are word embeddings more like definitions or spelling?**

Theoretical Background

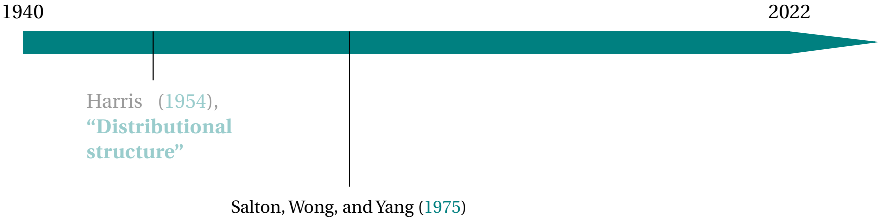
Chronology

1940 2022

Harris (1954), **"Distributional structure"**

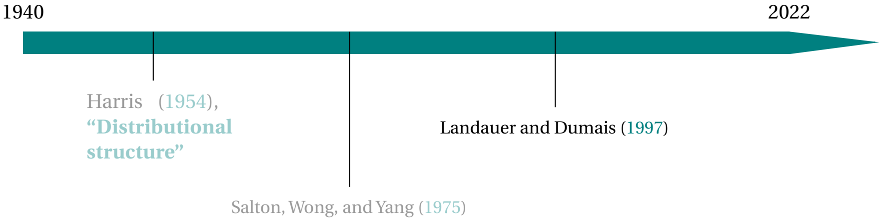
Seminal paper in Distributional Semantics

- Distributional Hypothesis (DH): Meaning should correlate with distribution



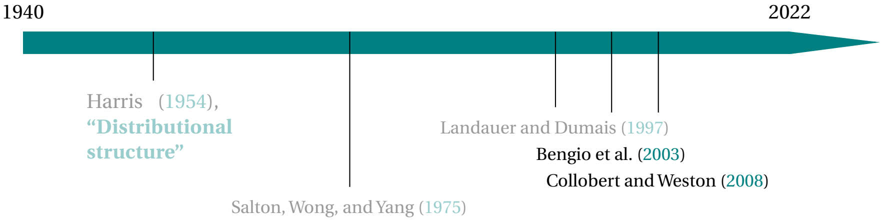
First large-scale vector model

- Designed for document vectors, not word vectors



First widely adopted Distributional Semantics Models (DSMs)

- Count-based models

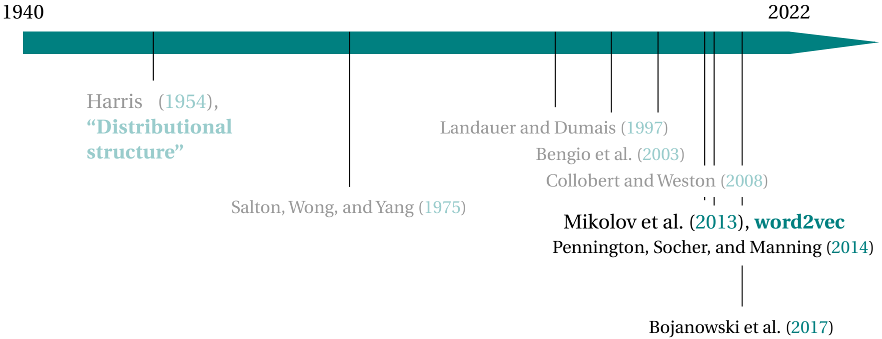


First neural word embeddings

- Bengio et al. (2003): Start of neural word embeddings
- Collobert and Weston (2008): Word embeddings as a multi-task framework

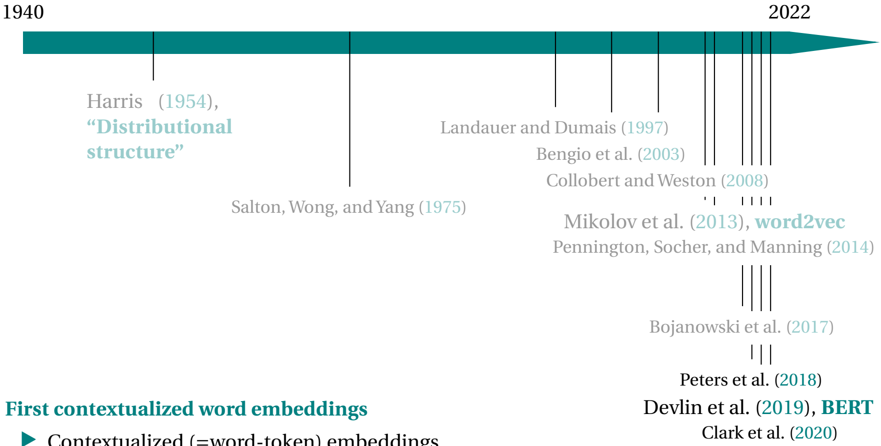
Theoretical Background

Chronology



Wide adoption of neural word embeddings

- Revolutionary
- Static (=word-type) representations
- Shallow neural network-based



First contextualized word embeddings

- Contextualized (=word-token) embeddings
- Often based on Transformer architecture (Vaswani et al., 2017)
- "One size fits all"

Distributional semantics models ≠ word embedding models

- Word embedding models are algorithms that convert words into vectors
- Distributional Semantics Models (DSMs) are meaningful vectors computed from distribution

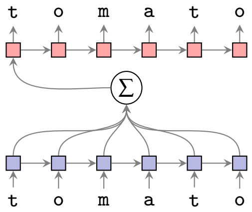
Theoretical Background

Different types of embeddings

Distributional semantics models ≠ word embedding models

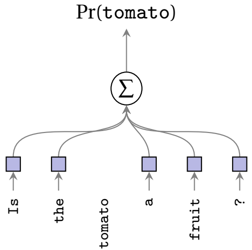
- Word embedding models are algorithms that convert words into vectors
- Distributional Semantics Models (DSMs) are meaningful vectors computed from distribution

Char-based



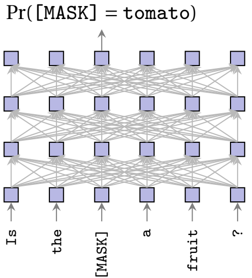
not distributional

word2vec



embeddings per word types

BERT



embeddings per word tokens

- How do word embeddings compare to dictionaries?

- How do word embeddings compare to dictionaries?

- First: what is a dictionary?

- How do word embeddings compare to dictionaries?

- First: what is a dictionary?

- Here:
  1. a dictionary is a list of definitions
  2. a definition links a **definiendum** to a *gloss*

- How do word embeddings compare to dictionaries?

- First: what is a dictionary?

- Here:
  1. a dictionary is a list of definitions
  2. a definition links a **definiendum** to a *gloss*

$$\text{Dict} = \left\{ \begin{array}{cc} \textbf{mirth} & \textit{The emotion usually following humour} \\ & \textit{and accompanied by laughter.} \\ \textbf{delight} & \textit{Joy; pleasure.} \\ \textbf{unquenched} & \textit{Not quenched.} \\ \dots & \dots \end{array} \right\}$$

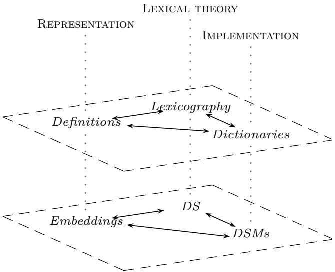
- How do word embeddings compare to dictionaries?
- First: what is a dictionary?
- Here:
  1. a dictionary is a list of definitions
  2. a definition links a **definiendum** to a *gloss*

$$\text{Dict} = \left\{ \begin{array}{cc} \textbf{mirth} & \textit{The emotion usually following humour} \\ & \textit{and accompanied by laughter.} \\ \textbf{delight} & \textit{Joy; pleasure.} \\ \textbf{unquenched} & \textit{Not quenched.} \\ \dots & \dots \end{array} \right\}$$

- Multiple patterns: Genus + Differentia, lists of near-synonyms, negated antonyms...

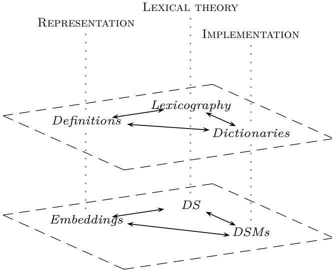
Theoretical Background

Side-by-side comparison



Theoretical Background

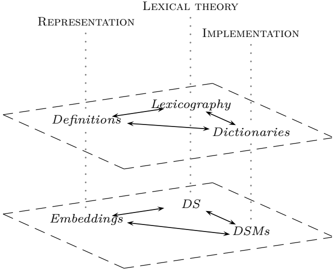
Side-by-side comparison



- Lexicography assumes language suffices to describe meaning
- DS assumes distribution suffices to describe meaning

Theoretical Background

Side-by-side comparison

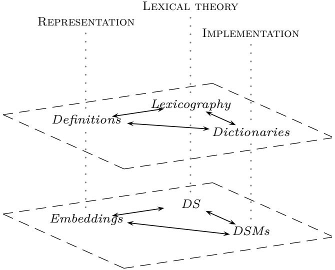


- Lexicography assumes language suffices to describe meaning
- Definitions are sequences of words

- DS assumes distribution suffices to describe meaning
- Embeddings are vectors

Theoretical Background

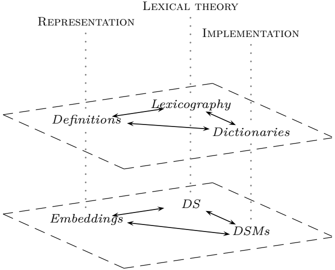
Side-by-side comparison



- Lexicography assumes language suffices to describe meaning
- Definitions are sequences of words
- Definitions are hand-crafted
- DS assumes distribution suffices to describe meaning
- Embeddings are vectors
- Embeddings are computed automatically

Theoretical Background

Side-by-side comparison



- Lexicography assumes language suffices to describe meaning
- Definitions are sequences of words
- Definitions are hand-crafted
- Different dictionaries make different assumptions about meaning

- DS assumes distribution suffices to describe meaning
- Embeddings are vectors
- Embeddings are computed automatically
- Different embedding models make different assumptions about meaning

Shopping list

To what extent are word embeddings lexical semantic representations?

Shopping list

To what extent are word embeddings lexical semantic representations?

1. Lexical semantic theories should be comparable

   If theory A says "ducks" and "geese" are similar, theory B shouldn't say they're unrelated

To what extent are word embeddings lexical semantic representations?

1. Lexical semantic theories should be comparable

   If theory A says "ducks" and "geese" are similar, theory B shouldn't say they're unrelated

2. Lexical semantic representations should be distinguishable from non-semantic ones

   We should be able to distinguish a definition from a string of random words

To what extent are word embeddings lexical semantic representations?

1. Lexical semantic theories should be comparable

   If theory A says "ducks" and "geese" are similar, theory B shouldn't say they're unrelated

2. Lexical semantic representations should be distinguishable from non-semantic ones

   We should be able to distinguish a definition from a string of random words

3. Lexical semantic representations should match predictions from their theory

   We don't want a definition for a word that says "this word can't be defined"

To what extent are word embeddings lexical semantic representations?

1. Lexical semantic theories should be comparable

   If theory A says "ducks" and "geese" are similar, theory B shouldn't say they're unrelated

2. Lexical semantic representations should be distinguishable from non-semantic ones

   We should be able to distinguish a definition from a string of random words

3. Lexical semantic representations should match predictions from their theory

   We don't want a definition for a word that says "this word can't be defined"

4. Lexical semantic representations should not encode non-semantic information

   Definitions need note include the price of the dictionary

- In our shopping list:
  1. Lexical semantic theories should be comparable

- In our shopping list:
  1. Lexical semantic theories should be comparable

- **How can we compare different types of representations such as vectors & sequences of words?**

- In our shopping list:
  1. Lexical semantic theories should be comparable

- **How can we compare different types of representations such as vectors & sequences of words?**

- Let's try to be exhaustive and look at multiple languages

  `en, es, fr, it, ru`

Experiments

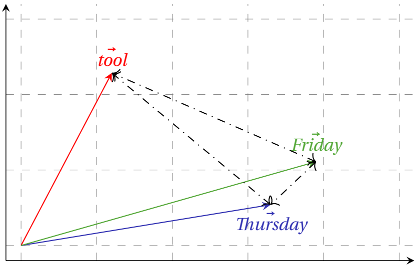
Comparing vectors & sequences

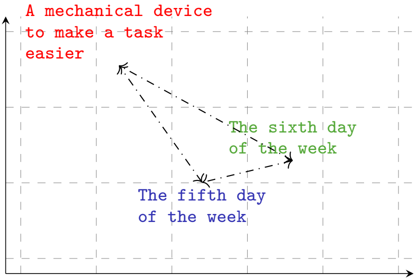
- We can rely on distances and use topographic similarity (Kirby, Cornish, and Smith, 2008) using a Mantel test

Experiments

Comparing vectors & sequences

- We can rely on distances and use topographic similarity (Kirby, Cornish, and Smith, 2008) using a Mantel test

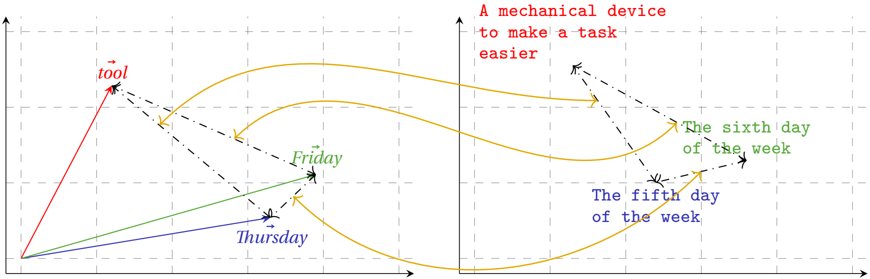




Experiments

Comparing vectors & sequences

- We can rely on distances and use topographic similarity (Kirby, Cornish, and Smith, 2008) using a Mantel test

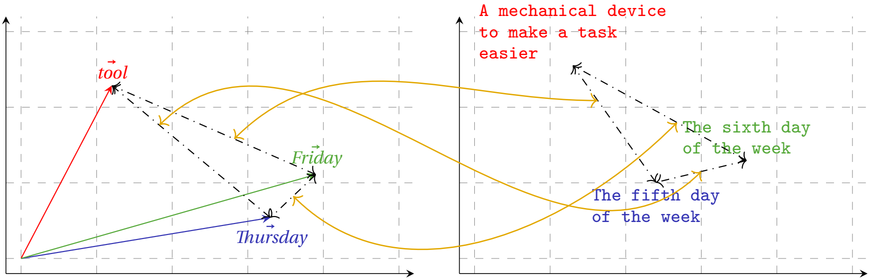


- We compute the correlation of all pairwise distance measurements

Experiments

Comparing vectors & sequences

- We can rely on distances and use topographic similarity (Kirby, Cornish, and Smith, 2008) using a Mantel test

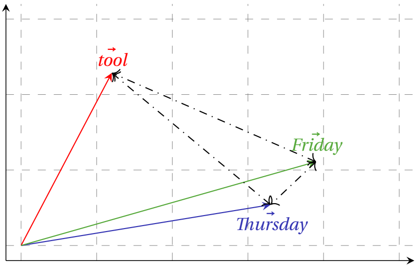


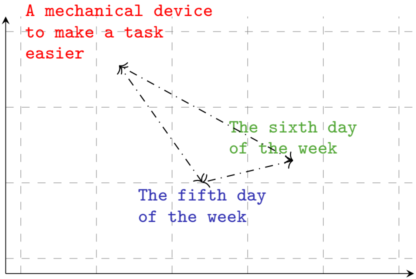
- We compute the correlation of all pairwise distance measurements
- Statistical significance is derived by comparing the observed correlation to random pairings

Experiments

Comparing vectors & sequences

- We can rely on distances and use topographic similarity (Kirby, Cornish, and Smith, 2008) using a Mantel test

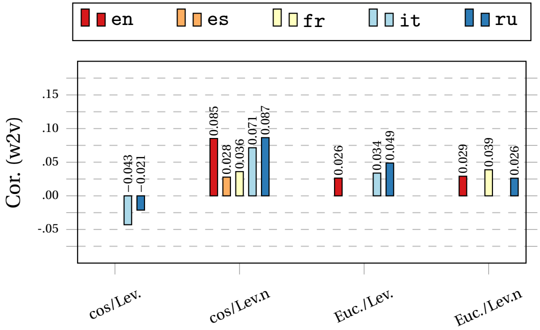




- We compute the correlation of all pairwise distance measurements
- Statistical significance is derived by comparing the observed correlation to random pairings
- Testing cosine & Euclidean distance for embeddings, and Levenshtein distance with or without normalization for definitions

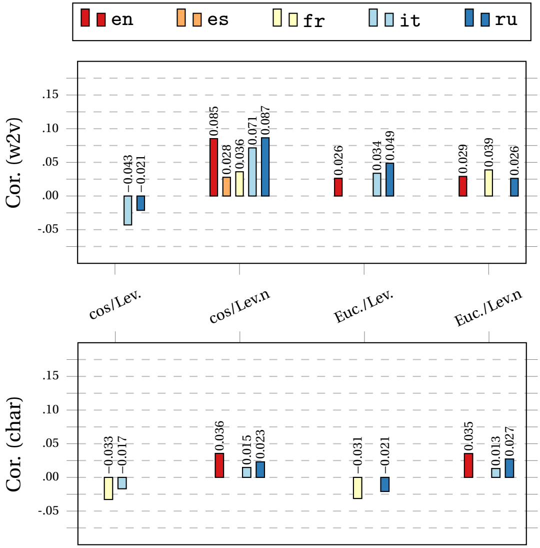
Experiments

What this looks like



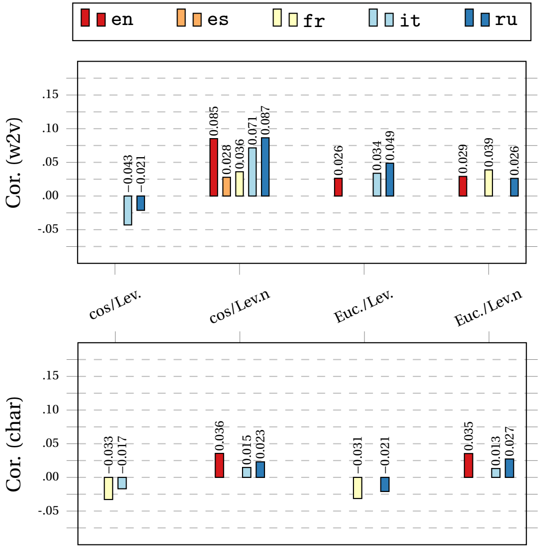
Experiments

What this looks like

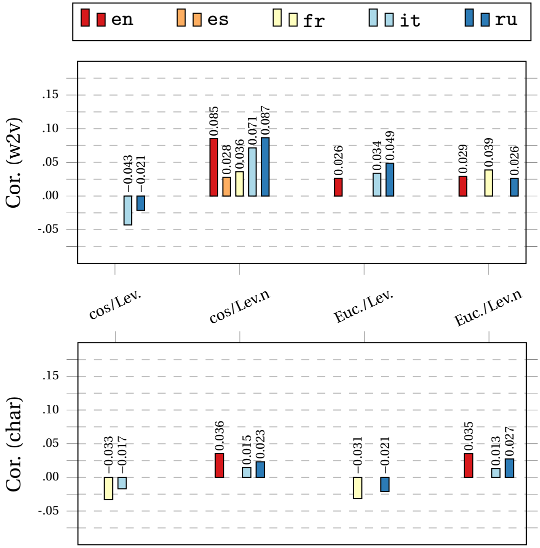


Experiments

What this looks like



As far as our shopping list is concerned:

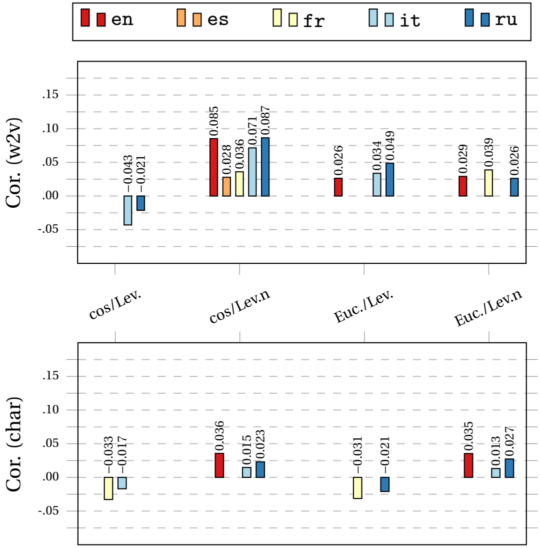


As far as our shopping list is concerned:

1. Lexical semantic theories should be comparable
   ⚠ **We find low correlations to low anti-correlations**

Experiments

What this looks like



As far as our shopping list is concerned:

1. Lexical semantic theories should be comparable
   ⚠ **We find low correlations to low anti-correlations**
2. Lexical semantic representations should be distinguishable from non-semantic ones
   ✓ **Character-based representations are worse than distributional ones**

Experiments

Pause for thoughts

- We could (and have) tested more complex metrics

- We could (and have) tested more complex metrics
- That would shift us from a non-parametric method to a parametric method

- We could (and have) tested more complex metrics
- That would shift us from a non-parametric method to a parametric method
- That would shift us from *measuring* a correlation to *modeling* a metric

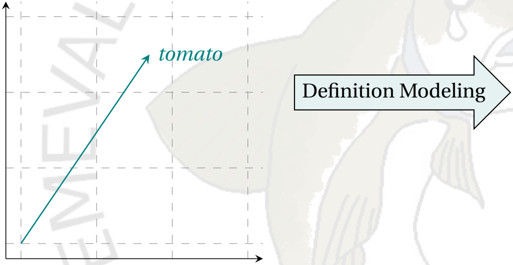
- We could (and have) tested more complex metrics
- That would shift us from a non-parametric method to a parametric method
- That would shift us from *measuring* a correlation to *modeling* a metric
- We might as well go all out: rather than modeling the metric, modeling the space

Experiments

As inverse functions

- Under a modeling perspective, we'd convert definitions into embeddings and back

- Under a modeling perspective, we'd convert definitions into embeddings and back



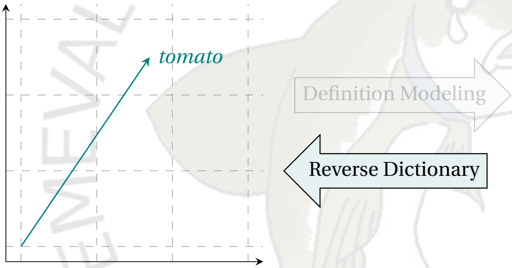
A widely cultivated plant, Solanum lycopersicum, having edible fruit.

Cf. Noraset et al. (2017)

Experiments

As inverse functions

- Under a modeling perspective, we'd convert definitions into embeddings and back



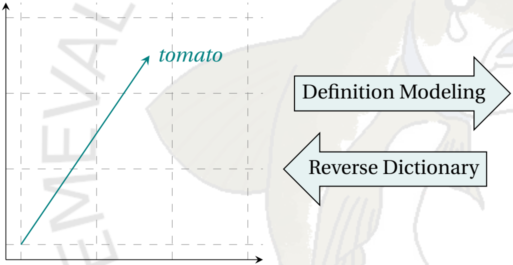
A widely cultivated plant, Solanum lycopersicum, having edible fruit.

Cf. Zanzotto et al. (2010), Hill et al. (2016)

Experiments

As inverse functions

- Under a modeling perspective, we'd convert definitions into embeddings and back



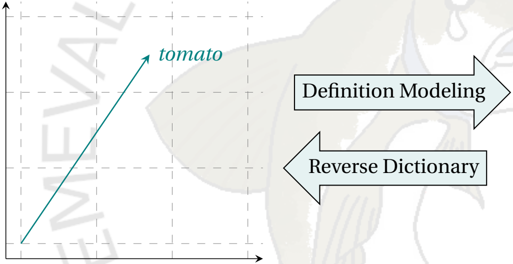
A widely cultivated plant, Solanum lycopersicum, having edible fruit.

- Shared task at SemEval 2022: CODWOE – Comparing Dictionaries and Word Embeddings
  159 valid submissions, 15+ different users, 11 system papers

Experiments

As inverse functions

- Under a modeling perspective, we'd convert definitions into embeddings and back



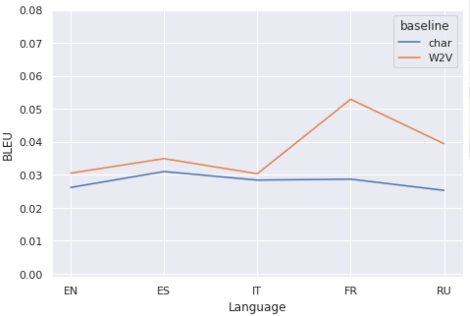
A widely cultivated plant, Solanum lycopersicum, having edible fruit.

- Shared task at SemEval 2022: CODWOE – Comparing Dictionaries and Word Embeddings
  159 valid submissions, 15+ different users, 11 system papers

- Focusing on DefMod BLEU results

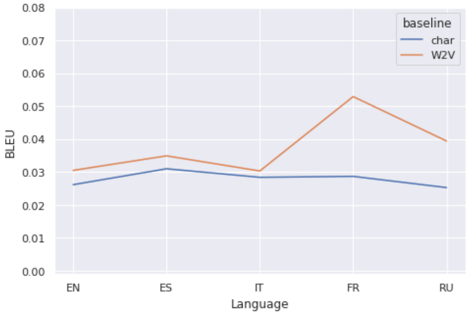
- Using simple LM baselines, seeded with definiendum embeddings

- Using simple LM baselines, seeded with definiendum embeddings



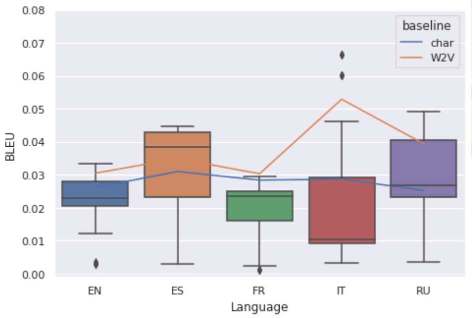
- ✓ **char embeddings rank systematically lower than W2V embeddings**

- Using simple LM baselines, seeded with definiendum embeddings



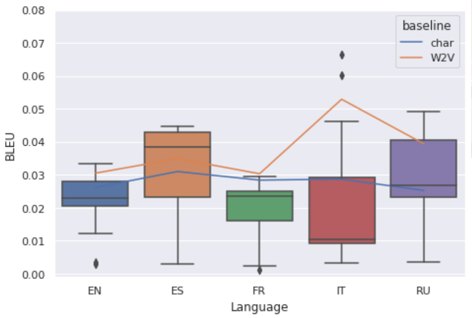
- ✓ **char embeddings rank systematically lower than W2V embeddings**
- ⚠ **Results are quantitatively low**
  Nonsensical outputs such as `", or ."` yield BLEU scores between 0.0189 and 0.0306 (Chen and Zhao, 2022)

- Using simple LM baselines, seeded with definiendum embeddings



- ✓ **char embeddings rank systematically lower than W2V embeddings**
- ⚠ **Results are quantitatively low**
  Nonsensical outputs such as `", or ."` yield BLEU scores between 0.0189 and 0.0306 (Chen and Zhao, 2022)
- Baselines are roughly in the middle of the submissions we received

- Using simple LM baselines, seeded with definiendum embeddings



- ✓ **char embeddings rank systematically lower than W2V embeddings**
- ⚠ **Results are quantitatively low**
  Nonsensical outputs such as `", or ."` yield BLEU scores between 0.0189 and 0.0306 (Chen and Zhao, 2022)
- Baselines are roughly in the middle of the submissions we received

- **Can we explain that?**

Why DefMod fails

Examples of usage

- **Word tokens & types do not necessarily coincide with word senses**
  Define "*tie*"

Why DefMod fails

Examples of usage

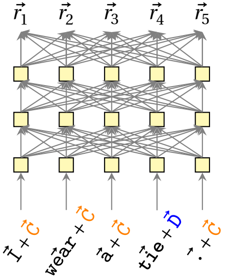
- **Word tokens & types do not necessarily coincide with word senses**
  Define “*tie*”

- This could be fixed by examples of usage:
  Define “*tie*” as in “I wear a tie.”

- **Word tokens & types do not necessarily coincide with word senses**
  Define "*tie*"

- This could be fixed by examples of usage:
  Define "*tie*" as in "I wear a tie."

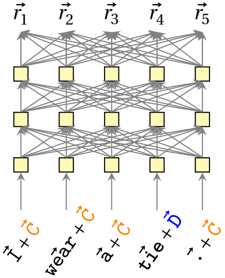
- Using sequence-to-sequence models



- **Word tokens & types do not necessarily coincide with word senses**
  Define "*tie*"

- This could be fixed by examples of usage:
  Define "*tie*" as in "I wear a tie."

- Using sequence-to-sequence models



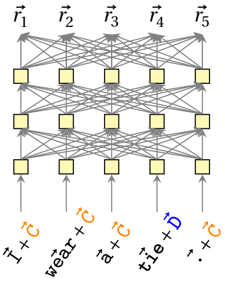
- Results in perplexity (how unlikely the productions are)
  with context: 33.6775
  without: 39.4279

  In line with the rest of the literature, e.g. Gadetsky, Yakubovskiy, and Vetrov (2018)

- **Word tokens & types do not necessarily coincide with word senses**
  Define "*tie*"

- This could be fixed by examples of usage:
  Define "*tie*" as in "I wear a tie."

- Using sequence-to-sequence models



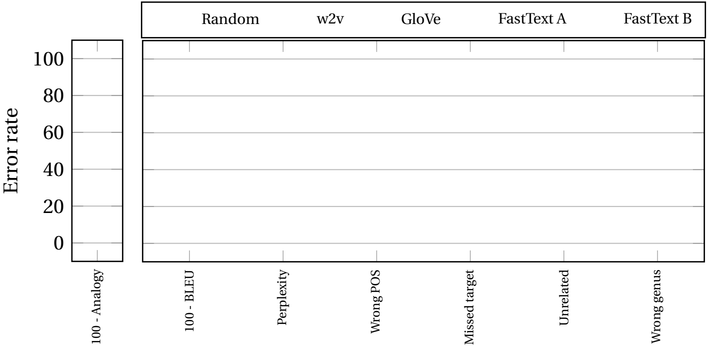
- Results in perplexity (how unlikely the productions are)
  with context: 33.6775
  without: 39.4279

  In line with the rest of the literature, e.g. Gadetsky, Yakubovskiy, and Vetrov (2018)

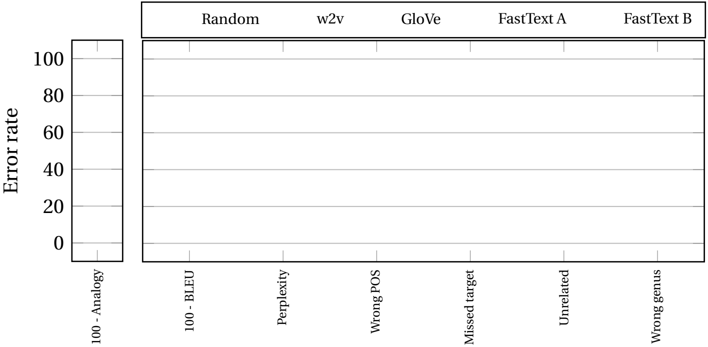
- ✗ **Definition Modeling can't work with embeddings alone**

- **Definition Modeling doesn't discriminate between embeddings**

- **Definition Modeling doesn't discriminate between embeddings**
- Let's compare sequence-to-sequence models trained on various embeddings with results on an analogy benchmark



- **Definition Modeling doesn't discriminate between embeddings**
- Let's compare sequence-to-sequence models trained on various embeddings with results on an analogy benchmark

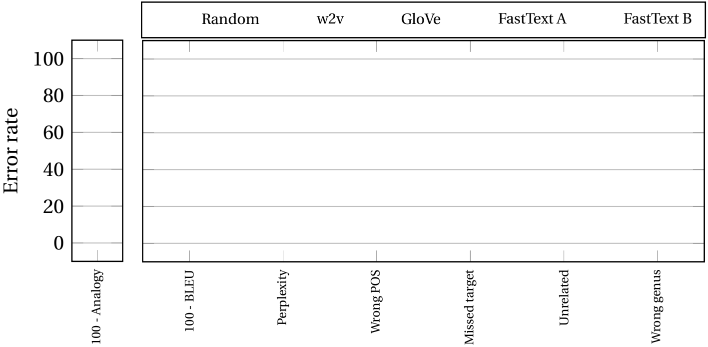


Wrong POS:

les rives de l'Orange offraient toujours le même aspect ***enchanteur***

Enchanteur: personne qui rêve

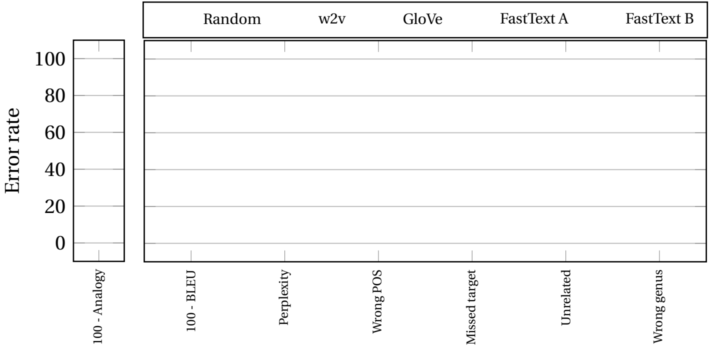
- **Definition Modeling doesn't discriminate between embeddings**
- Let's compare sequence-to-sequence models trained on various embeddings with results on an analogy benchmark



Missed target:

*Elle venait de créer ce qu'on nommait des **bons** de délégation ...*
Bon: qui est bon, heureux favorable

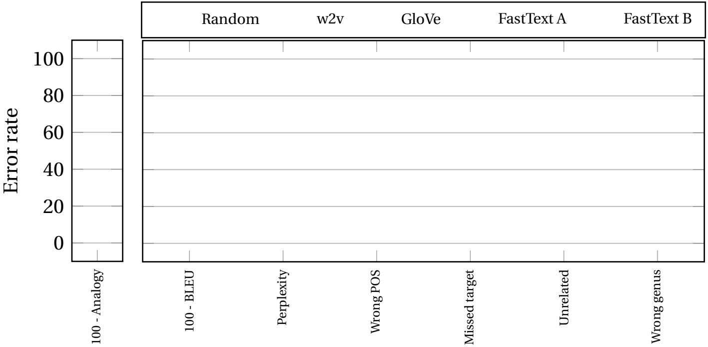
- **Definition Modeling doesn't discriminate between embeddings**
- Let's compare sequence-to-sequence models trained on various embeddings with results on an analogy benchmark



Unrelated:

Chercheur: étoffe de soie, de coton, etc.

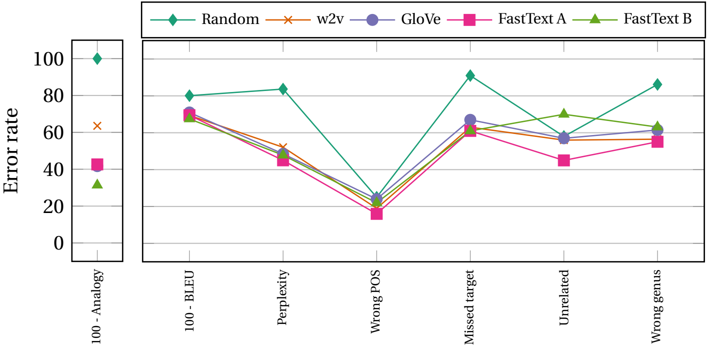
- **Definition Modeling doesn't discriminate between embeddings**
- Let's compare sequence-to-sequence models trained on various embeddings with results on an analogy benchmark



Wrong genus:

Kilomole: anion de bismuth

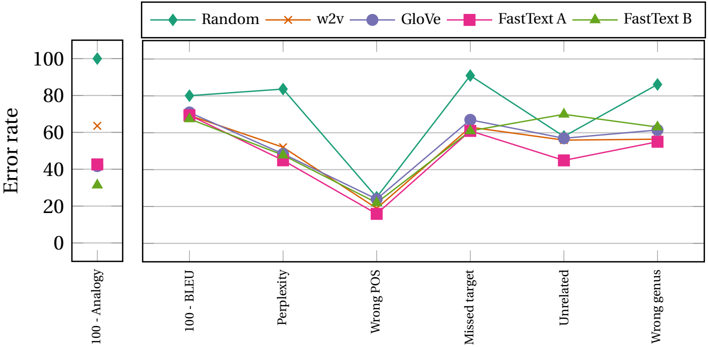
- **Definition Modeling doesn't discriminate between embeddings**
- Let's compare sequence-to-sequence models trained on various embeddings with results on an analogy benchmark



Why DefMod fails

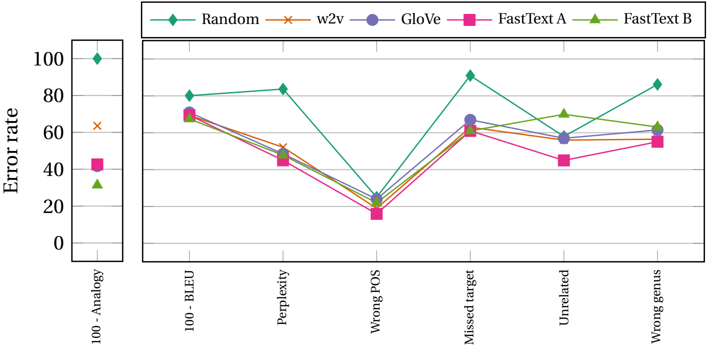
Quality of embeddings

- **Definition Modeling doesn't discriminate between embeddings**
- Let's compare sequence-to-sequence models trained on various embeddings with results on an analogy benchmark



- ✓ **DefMod distinguishes random & trained embeddings**

- **Definition Modeling doesn't discriminate between embeddings**
- Let's compare sequence-to-sequence models trained on various embeddings with results on an analogy benchmark



- ✓ **DefMod distinguishes random & trained embeddings**
- ✗ **Unlike analogy, DefMod doesn't clearly distinguish between embeddings**

- Back to our shopping list:
  1. Lexical semantic theories should be comparable
     ⚠ **We get at best a low correlation between embeddings & definition spaces**

     ✗ **Word embeddings do not coincide with word senses**
  2. Lexical semantic representations should be distinguishable from non-semantic ones
     ✓ **We do distinguish char-based & random embeddings from distributional embeddings**

- Next up on the list:
  3. Lexical semantic representations should match predictions from their theory

- Next up on the list:
  3. Lexical semantic representations should match predictions from their theory

- **Let's have a look at Harris (1954)**

> *Substitutability (parallel). It will in general appear that various elements have identical types of occurrence-dependence. We group A and B into a substitution set whenever A and B each have the same (or partially same) environments X (X being at first elements, later substitution sets of elements) within a statable domain of the flow of speech. This enables us to speak of the occurrence-dependence of a whole set of elements in respect to other such sets of elements.*

Harris (1954)

> *Substitutability (parallel). It will in general appear that various elements have identical types of occurrence-dependence. We group A and B into a substitution set whenever A and B each have the same (or partially same) environments X (X being at first elements, later substitution sets of elements) within a statable domain of the flow of speech. This enables us to speak of the occurrence-dependence of a whole set of elements in respect to other such sets of elements.*

Harris (1954)

- "*Friday*" and "*Thursday*" should be substitutable with one another, but not with "*tool*"

Distributional Substitutability

What we expect of DSMs

> *Substitutability (parallel). It will in general appear that various elements have identical types of occurrence-dependence. We group A and B into a substitution set whenever A and B each have the same (or partially same) environments X (X being at first elements, later substitution sets of elements) within a statable domain of the flow of speech. This enables us to speak of the occurrence-dependence of a whole set of elements in respect to other such sets of elements.*

Harris (1954)

- "*Friday*" and "*Thursday*" should be substitutable with one another, but not with "*tool*"
- We can tweak it to test embedding algorithms:

$$\Pr(w_1|c) > \Pr(w_2|c)$$

  For substitutable words, this difference should be small, and large otherwise

> *Substitutability (parallel). It will in general appear that various elements have identical types of occurrence-dependence. We group A and B into a substitution set whenever A and B each have the same (or partially same) environments X (X being at first elements, later substitution sets of elements) within a statable domain of the flow of speech. This enables us to speak of the occurrence-dependence of a whole set of elements in respect to other such sets of elements.*
>
> Harris (1954)

- "*Friday*" and "*Thursday*" should be substitutable with one another, but not with "*tool*"
- We can tweak it to test embedding algorithms:

  $$\Pr(w_1|c) > \Pr(w_2|c)$$

  For substitutable words, this difference should be small, and large otherwise
- We can compare human intuitions to word embedding predictions

- Basically a fill-in-the-gaps test:

- Basically a fill-in-the-gaps test:

```
best way to dissect the aortic
______.
the ______ and pericardium have
both been recorded as points of
outlet.
if the ______ be implicated,
greater expansion of the upper
and outside portion of the left
side of the chest in inspiration
takes place.
```

pleura? *diaphragm*? *elevator*?

- Basically a fill-in-the-gaps test:

  ```
  best way to dissect the aortic
  _______.
  the _______ and pericardium have
  both been recorded as points of
  outlet.
  if the _______ be implicated,
  greater expansion of the upper
  and outside portion of the left
  side of the chest in inspiration
  takes place.
  ```

  pleura? *diaphragm*? *elevator*?

- We can turn this into an online game:
  `https://blankcrack.atilf.fr`

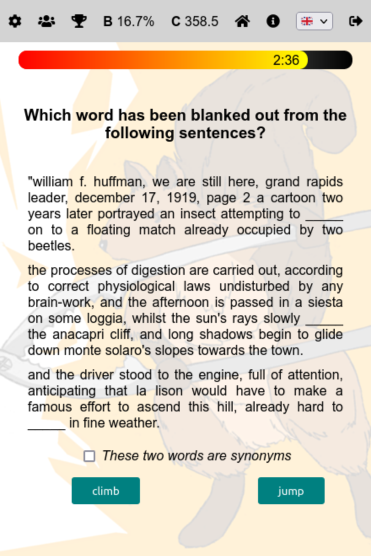
BlankCrack

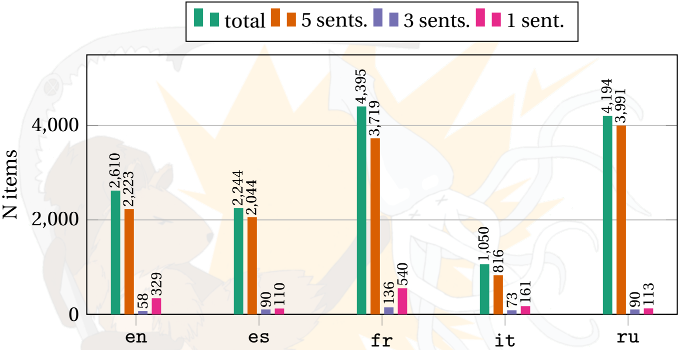
- Basically a fill-in-the-gaps test:

```
best way to dissect the aortic
______.
the ______ and pericardium have
both been recorded as points of
outlet.
if the ______ be implicated,
greater expansion of the upper
and outside portion of the left
side of the chest in inspiration
takes place.
```

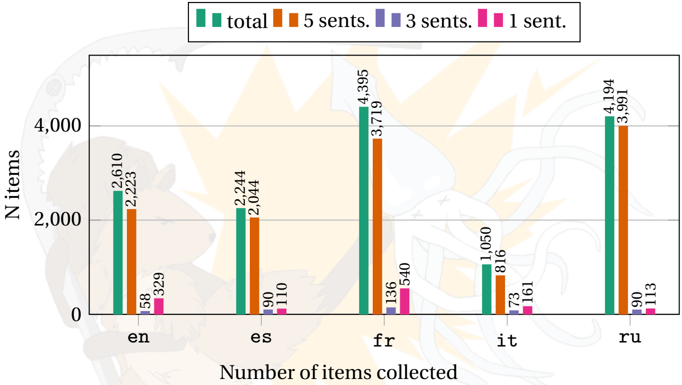
  pleura? *diaphragm*? *elevator*?

- We can turn this into an online game: `https://blankcrack.atilf.fr`

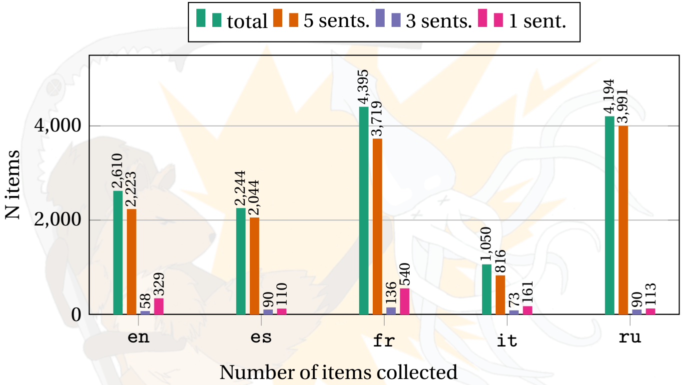




Number of items collected



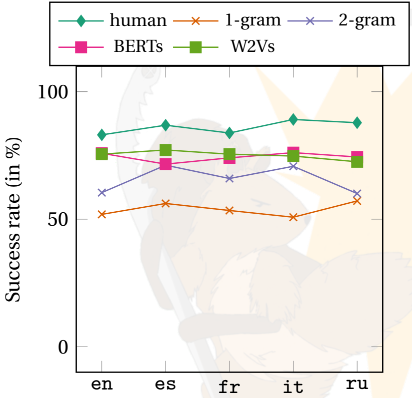
- ⚠ **Still a small dataset**



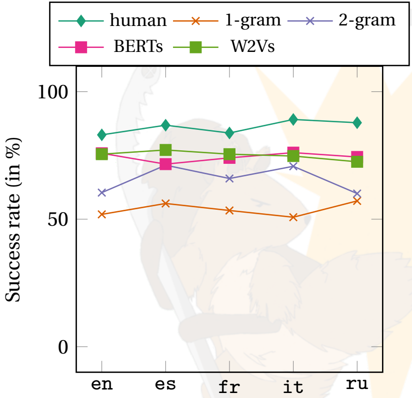
- ⚠ **Still a small dataset**
- ✓ **Some very hard pairs**
  `baseball` vs. `basketball`, `aquarelle` vs. `gouache`...

How do human intuitions compare to word embedding predictions?

How do human intuitions compare to word embedding predictions?

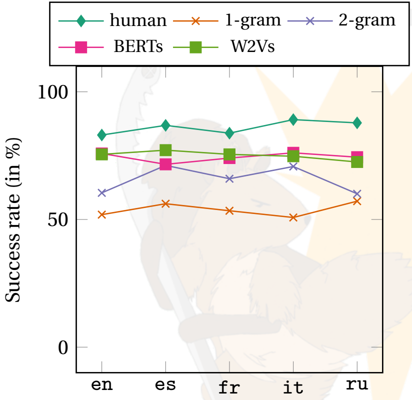


How do human intuitions compare to word embedding predictions?



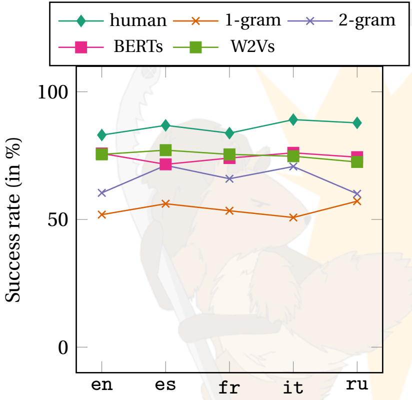
- ✓ **Embeddings perform better than n-grams**

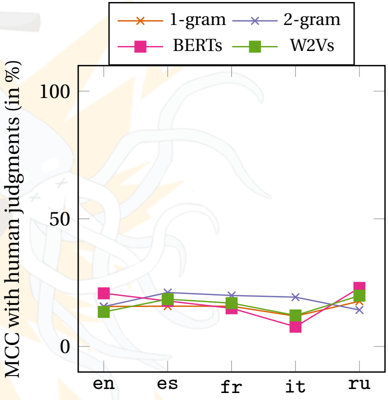
How do human intuitions compare to word embedding predictions?



- ✓ **Embeddings perform better than n-grams**
- ⚠ **Noticeable gap with human performance**

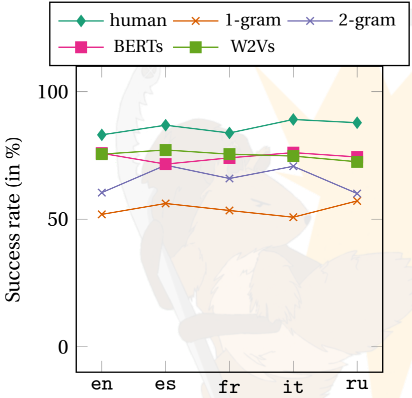
How do human intuitions compare to word embedding predictions?

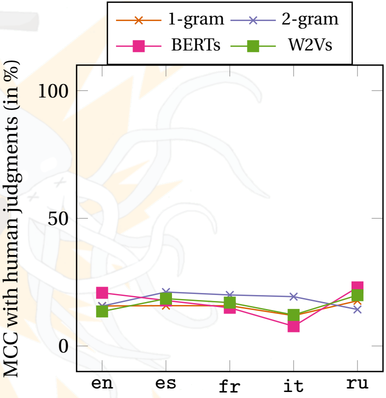




- ✓ Embeddings perform better than n-grams
- ⚠ Noticeable gap with human performance

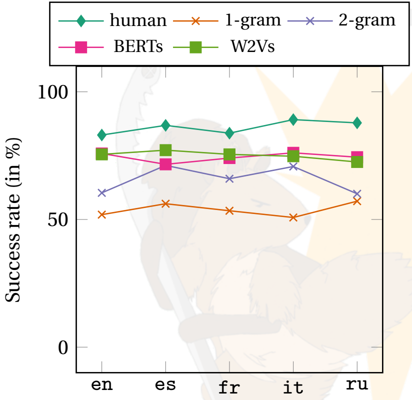
How do human intuitions compare to word embedding predictions?

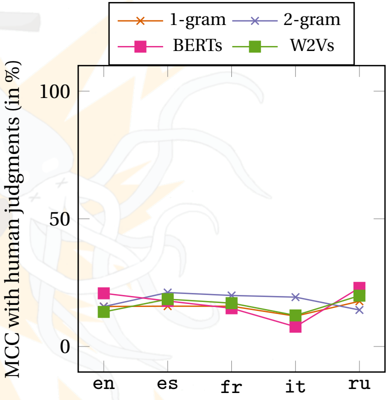




- ✓ **Embeddings perform better than n-grams**
- ⚠ **Noticeable gap with human performance**
- ✓ **Positive correlation with human behavior**

How do human intuitions compare to word embedding predictions?





- ✓ Embeddings perform better than n-grams
- ⚠ Noticeable gap with human performance

- ✓ Positive correlation with human behavior
- ✗ Embeddings do not contrast with n-grams

- We were looking at
  3. Lexical semantic representations should match predictions from their theory
- ✗ **Embeddings underperfom humans on substitutability judgments**
- ✗ **Embeddings do not model human behavior any better than n-grams**

Experiments

Second pause for thoughts

- We were looking at
  3. Lexical semantic representations should match predictions from their theory
- ✗ **Embeddings underperfom humans on substitutability judgments**
- ✗ **Embeddings do not model human behavior any better than n-grams**
- Embeddings nonetheless perform decently

- We were looking at
  3. Lexical semantic representations should match predictions from their theory
- **✗ Embeddings underperfom humans on substitutability judgments**
- **✗ Embeddings do not model human behavior any better than n-grams**
- Embeddings nonetheless perform decently
- **Should we instead analyze their behavior algorithmically?**
  i.e., check
  4. Lexical semantic representations should not encode non-semantic information

Muppet Dissection

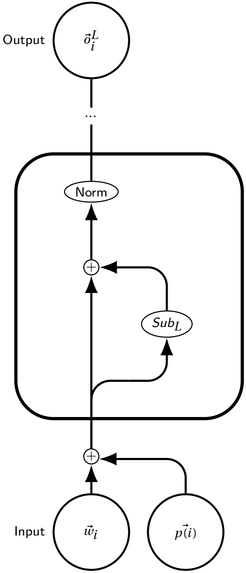
Sentence bias

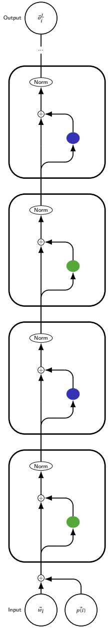
- We focus on BERT

Muppet Dissection

Sentence bias

- We focus on BERT
- BERT is a Transformer
  - a stack of sublayers
  - **multihead attention** / **feed-forwards** sublayer functions
  - vector inputs
  - layer-normalizations & residual connections

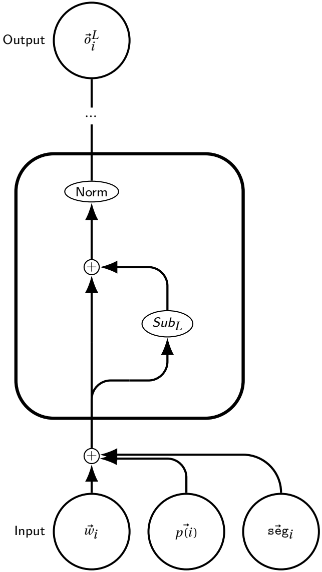


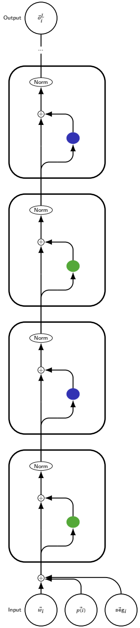


Muppet Dissection

Sentence bias

- We focus on BERT
- BERT is a Transformer
  - a stack of sublayers
  - **multihead attention** / **feed-forwards** sublayer functions
  - vector inputs
  - layer-normalizations & residual connections
- Relies on a MLM objective
  $\Pr(\texttt{[MASK]} = w|c)$
  and a NSP objective:
  $\Pr(S_A \prec S_B | S_A, S_B)$
  with $\vec{\text{seg}}_A, \vec{\text{seg}}_B$ to distinguish S_A, S_B

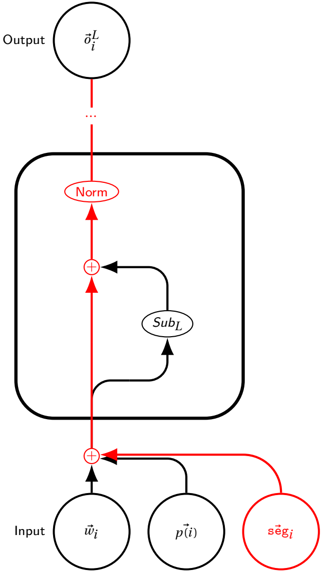


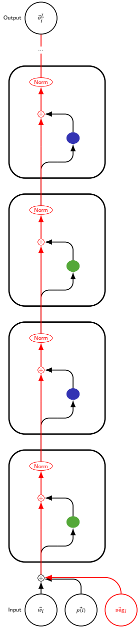


Muppet Dissection

Sentence bias

- We focus on BERT
- BERT is a Transformer
  - a stack of sublayers
  - **multihead attention** / **feed-forwards** sublayer functions
  - vector inputs
  - layer-normalizations & residual connections
- Relies on a MLM objective
  $\Pr(\texttt{[MASK]} = w|c)$
  and a NSP objective:
  $\Pr(S_A \prec S_B | S_A, S_B)$
  with $\vec{\text{seg}}_A, \vec{\text{seg}}_B$ to distinguish S_A, S_B
- Residual connections create a pathway





- The residual pathway means vector inputs bear a trace on the output

Muppet Dissection

Sentence bias

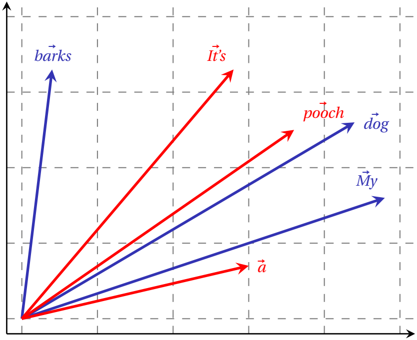
- The residual pathway means vector inputs bear a trace on the output
- Each embedding is shifted by a scaled segment encoding

Muppet Dissection

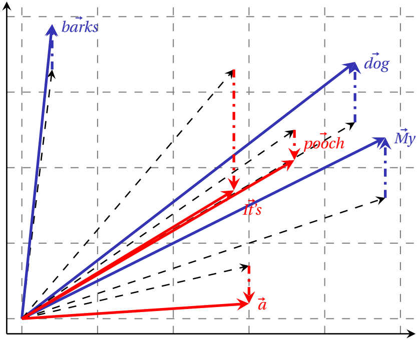
Sentence bias

- The residual pathway means vector inputs bear a trace on the output
- Each embedding is shifted by a scaled segment encoding

E.g., for the vectors: BERT(“*My dog barks.* *It’s a pooch.*”)



Toy example without bias

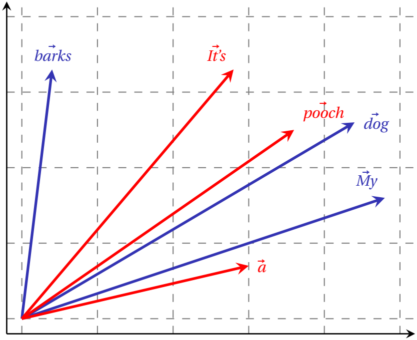


Toy example with bias

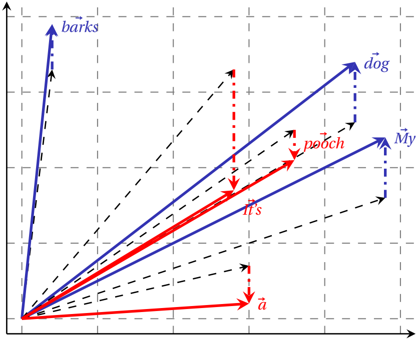
Muppet Dissection

- The residual pathway means vector inputs bear a trace on the output
- Each embedding is shifted by a scaled segment encoding

E.g., for the vectors: BERT("*My dog barks. It's a pooch.*")



Toy example without bias



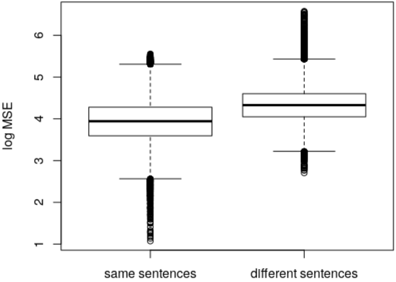
Toy example with bias

- **Is this bias noticeable?**

- Let's measure whether there's a noticeable difference between embeddings of the same type but different segments

  two occurences of "*tie*" in the same segment vs. two occurences of "*tie*" in different segments

- Let's measure whether there's a noticeable difference between embeddings of the same type but different segments
  two occurences of "*tie*" in the same segment vs. two occurences of "*tie*" in different segments

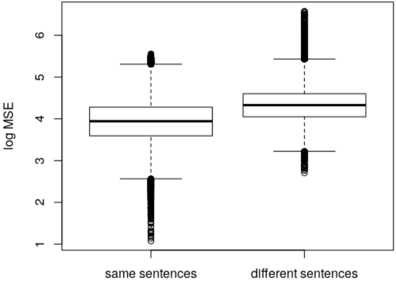


Muppet Dissection

Sentence bias

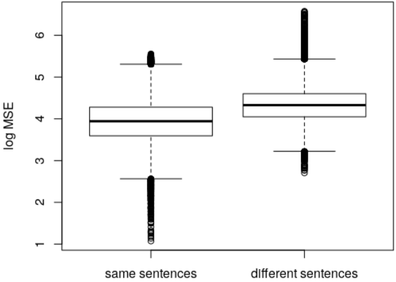
- Let's measure whether there's a noticeable difference between embeddings of the same type but different segments

  two occurences of "*tie*" in the same segment vs. two occurences of "*tie*" in different segments



- MSE scores systematically favor the mean of the token's own segment

- Let's measure whether there's a noticeable difference between embeddings of the same type but different segments
  two occurences of "*tie*" in the same segment vs. two occurences of "*tie*" in different segments



- MSE scores systematically favor the mean of the token's own segment
- Wrt. our last shopping list item:
  4. Lexical semantic representations should not encode non-semantic information

✗ **This bias is noticeable**

Muppet Dissection

But wait, it generalizes!

- The residual pathway means the output is a sum of sub-vectors
- We can decompose transformer embeddings in four terms: $\vec{e}_t = \vec{I} + \vec{F} + \vec{H} + \vec{C}$

Muppet Dissection

But wait, it generalizes!

- The residual pathway means the output is a sum of sub-vectors
- We can decompose transformer embeddings in four terms: $\vec{e}_t = \vec{I} + \vec{F} + \vec{H} + \vec{C}$
    1. a term related to the input, $\vec{I}$

Muppet Dissection

But wait, it generalizes!

- The residual pathway means the output is a sum of sub-vectors
- We can decompose transformer embeddings in four terms: $\vec{e}_t = \vec{I} + \vec{F} + \vec{H} + \vec{C}$

  1. a term related to the input, $\vec{I}$
  2. a term related to the feed-forward modules, $\vec{F}$
  3. a term related to the multihead attentions, $\vec{H}$

Muppet Dissection

But wait, it generalizes!

- The residual pathway means the output is a sum of sub-vectors
- We can decompose transformer embeddings in four terms: $\vec{e}_t = \vec{I} + \vec{F} + \vec{H} + \vec{C}$

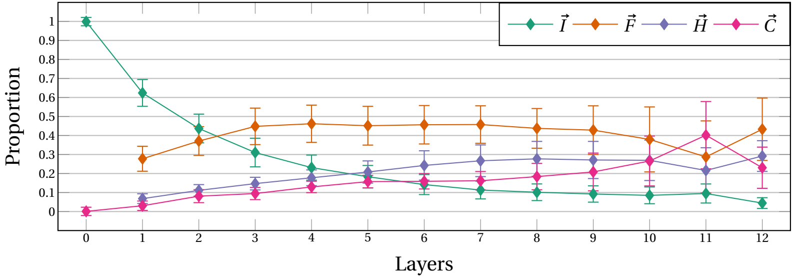
  1. a term related to the input, $\vec{I}$
  2. a term related to the feed-forward modules, $\vec{F}$
  3. a term related to the multihead attentions, $\vec{H}$
  4. a term where we collect biases and offsets, $\vec{C}$

- The residual pathway means the output is a sum of sub-vectors
- We can decompose transformer embeddings in four terms: $\vec{e}_t = \vec{I} + \vec{F} + \vec{H} + \vec{C}$

  1. a term related to the input, $\vec{I}$
  2. a term related to the feed-forward modules, $\vec{F}$
  3. a term related to the multihead attentions, $\vec{H}$
  4. a term where we collect biases and offsets, $\vec{C}$

- We can visualize the proportion of the embedding $\vec{e}_t$ corresponding to these four terms:

- The residual pathway means the output is a sum of sub-vectors
- We can decompose transformer embeddings in four terms: $\vec{e}_t = \vec{I} + \vec{F} + \vec{H} + \vec{C}$
  1. a term related to the input, $\vec{I}$
  2. a term related to the feed-forward modules, $\vec{F}$
  3. a term related to the multihead attentions, $\vec{H}$
  4. a term where we collect biases and offsets, $\vec{C}$
- We can visualize the proportion of the embedding $\vec{e}_t$ corresponding to these four terms:



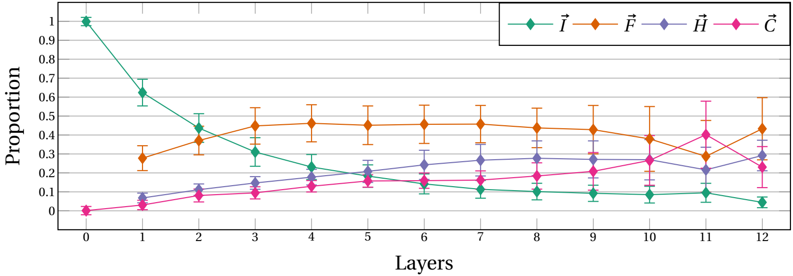
Muppet Dissection

But wait, it generalizes!

- The residual pathway means the output is a sum of sub-vectors
- We can decompose transformer embeddings in four terms: $\vec{e}_t = \vec{I} + \vec{F} + \vec{H} + \vec{C}$

  1. a term related to the input, $\vec{I}$
  2. a term related to the feed-forward modules, $\vec{F}$
  3. a term related to the multihead attentions, $\vec{H}$
  4. a term where we collect biases and offsets, $\vec{C}$

- We can visualize the proportion of the embedding $\vec{e}_t$ corresponding to these four terms:



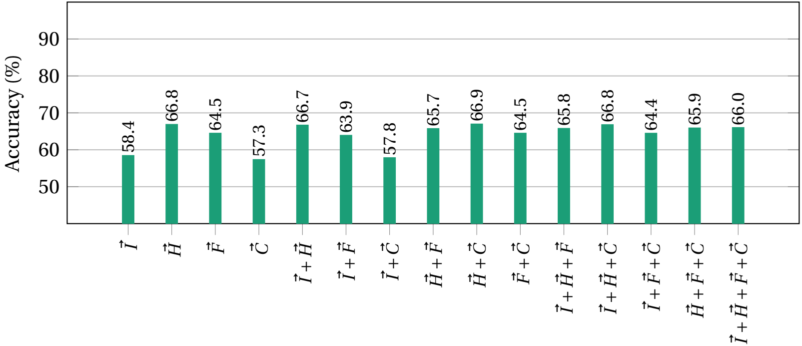
- **Do these different terms model lexical semantics differently?**

- Using WSD: lexical semantic representations should encode word senses

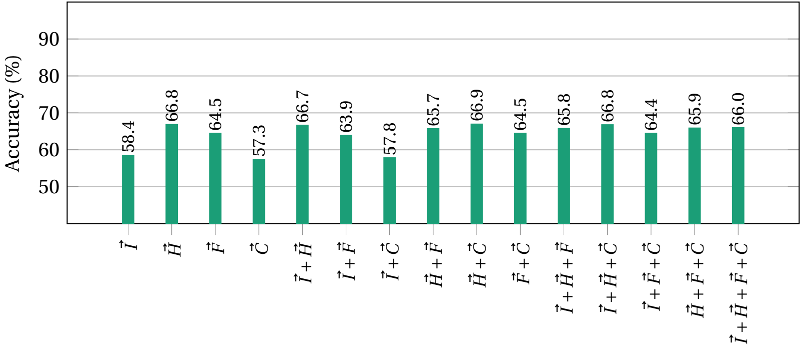
Muppet Dissection

Word Sense Disambiguation

- Using WSD: lexical semantic representations should encode word senses

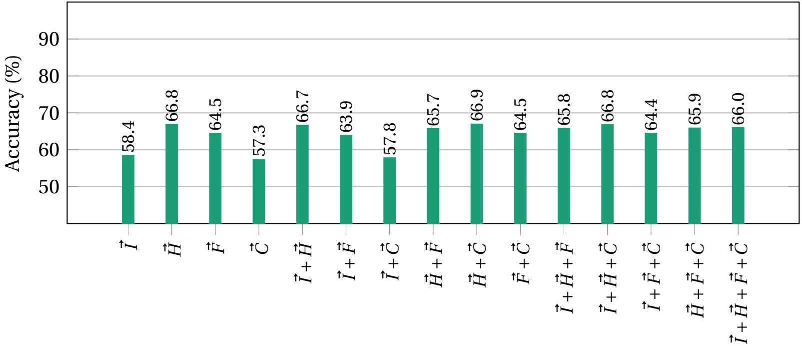


- Using WSD: lexical semantic representations should encode word senses



- The different terms all yield different results

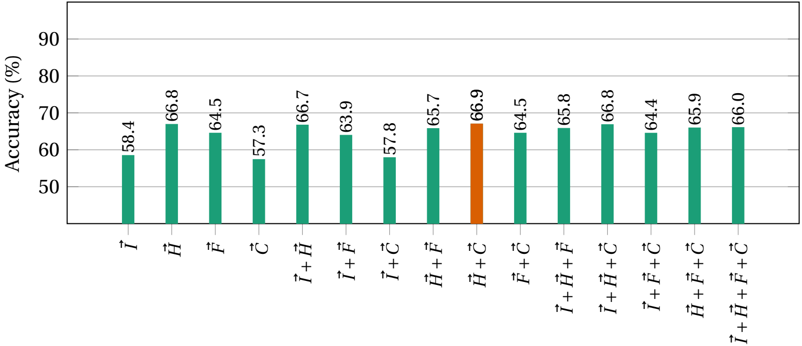
- Using WSD: lexical semantic representations should encode word senses



- The different terms all yield different results

- The full embedding isn't the one that performs best

- Using WSD: lexical semantic representations should encode word senses



- The different terms all yield different results
- The full embedding isn't the one that performs best

- When looking at:
  4. Lexical semantic representations should not encode non-semantic information

✗ There are obvious biases in Transformer embeddings due to their implementations
✗ These biases impact the quality of the overall embedding

Conclusions

To what extent are word embeddings lexical semantic representations?

Conclusions

To what extent are word embeddings lexical semantic representations?

1. Lexical semantic theories should be comparable
   ⚠ **We get at best a low correlation between embeddings & definition spaces**
   ✗ **We have an alignment problem**
2. Lexical semantic representations should be distinguishable from non-semantic ones
   ✓ **Char-based & random embeddings are distinct from distributional ones**
3. Lexical semantic representations should match predictions from their theory
   ✗ **Embeddings don't match our expectations for distributional substitutability**
4. Lexical semantic representations should not encode non-semantic information
   ✗ **We find obvious detrimental biases due to embedding implementation**

Conclusions

To what extent are word embeddings lexical semantic representations?

1. Lexical semantic theories should be comparable
   ⚠ **We get at best a low correlation between embeddings & definition spaces**
   ✗ **We have an alignment problem**
2. Lexical semantic representations should be distinguishable from non-semantic ones
   ✓ **Char-based & random embeddings are distinct from distributional ones**
3. Lexical semantic representations should match predictions from their theory
   ✗ **Embeddings don't match our expectations for distributional substitutability**
4. Lexical semantic representations should not encode non-semantic information
   ✗ **We find obvious detrimental biases due to embedding implementation**

In a nutshell:

- We can make quantitative statements about the fitness of DSMs as a semantic theory of the lexicon
- We should be more cautious about how we talk about DSMs and word embeddings

Thanks for your attention!

List of Publications

Mickus, Timothee, Timothée Bernard, and Denis Paperno (Dec. 2020). "What Meaning-Form Correlation Has to Compose With: A Study of MFC on Artificial and Natural Language".
Mickus, Timothee, Mathieu Constant, and Denis Paperno (June 2020). "Génération automatique de définitions pour le français (Definition Modeling in French)".
— (July 2021a). "A Game Interface to Study Semantic Grounding in Text-Based Models".
— (Dec. 2021b). "About Neural Networks and Writing Definitions".
Mickus, Timothee, Denis Paperno, and Mathieu Constant (Sept. 2019). "Mark my Word: A Sequence-to-Sequence Approach to Definition Modeling".
Mickus, Timothee et al. (Jan. 2020). "What do you mean, BERT?"

References

Bengio, Yoshua et al. (Mar. 2003). "A Neural Probabilistic Language Model".
Bojanowski, Piotr et al. (2017). "Enriching Word Vectors with Subword Information".
Chen, Pinzhen and Zheng Zhao (2022). *Edinburgh at SemEval-2022 Task 1: Jointly Fishing for Word Embeddings and Definitions*.
Clark, Kevin et al. (Nov. 2020). "Pre-Training Transformers as Energy-Based Cloze Models".
Collobert, Ronan and Jason Weston (2008). "A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning".
Devlin, Jacob et al. (June 2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding".
Gadetsky, Artyom, Ilya Yakubovskiy, and Dmitry Vetrov (2018). "Conditional Generators of Words Definitions".
Harris, Zellig (1954). "Distributional structure".
Hill, Felix et al. (2016). "Learning to Understand Phrases by Embedding the Dictionary".
Kirby, Simon, Hannah Cornish, and Kenny Smith (2008). "Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language".
Landauer, Thomas K and Susan T. Dumais (1997). "A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge".
Mikolov, Tomas et al. (Jan. 2013). "Efficient Estimation of Word Representations in Vector Space".
Noraset, Thanapon et al. (2017). "Definition Modeling: Learning to define word embeddings in natural language".
Pennington, Jeffrey, Richard Socher, and Christopher Manning (Oct. 2014). "GloVe: Global Vectors for Word Representation".
Peters, Matthew et al. (June 2018). "Deep Contextualized Word Representations".
Salton, G., A. Wong, and C. S. Yang (Nov. 1975). "A Vector Space Model for Automatic Indexing".
Vaswani, Ashish et al. (2017). "Attention is All you Need".
Zanzotto, Fabio Massimo et al. (Aug. 2010). "Estimating Linear Models for Compositional Distributional Semantics".