To the Limits of Distributional Semantics and Beyond

Denis Paperno

15.12.2022

Common assumptions

- Text is a limited source of information
- Perceptual grounding, typically visual, necessary for ‘real’ language understanding

Argument

- Text can go a very long way as a source of meaning representations
- Visual grounding is of limited relevance
- Knowledge about referential properties, e.g. from databases, is an important source of grounding. Example: geographic information

Distributional Semantics

Representations for language are learned from word distributions

- LSA, word2vec, GloVe, fastText
- more recently: pretrained large language models (BERT, GPT, etc.)

"impressive natural language understanding and generation capabilities"

(PaLM, Chowderry 2022)

Distributional Hypothesis

- Meaning distinctions are reflected in distributional differences

“It may be presumed that any two morphemes A and B having different meanings, also differ somewhere in distribution: there are some environments in which one occurs and the other does not”

(Harris 1951)

"You shall know a word by the company it keeps." (Firth 1957)

Example: what word is masked as XXXXX?

Abul-Hassan, the merchant’s son, on being shown the portrait of the lady, requested his father to delay the XXXXX till he could reconcile his mind to it.

In East Friesland, it is believed, when seven girls succeed each other in one family, that among them one is of necessity a were-wolf, so that youths are slow in seeking one of seven sisters in XXXXX.

According to a Polish story, if a witch lays a girdle of human skin on the threshold of a house in which a XXXXX is being celebrated, the bride and bridegroom, and bridesmaids and groomsmen, should they step across it, are transformed into wolves.

Distributional vectors

- Embeddings often encode co-occurrence properties of words

- A common idea:
  - find vectors of words w_i (e.g. *dog*) and contexts $\widetilde{w}_k$ (e.g. *bark*)
  - such that dot products of associated pairs $w_i \cdot \widetilde{w}_k$ is high
  - and for random pairs $w_i \cdot \widetilde{w}'_n$ is low (e.g. *dog vs. logarithm*)

  - often using learning techniques as in neural models (**skipgram**, **GloVe**)
  - Language models: larger contexts, >1 word (**CBOW**, **ELMo**, **BERT**, **GPT**…)

  - Learned word vectors can be compared for similarity

Relatedness/similarity evaluation

- Words with similar distributional vectors have related meanings

money vs. cash, .98 cosine

vs

stock vs. phone, .04 cosine

Example from WordSim353 (Finkelstein et al. 2002);

cosines from a word2vec model

Similarity is not all you need

- Words can have very similar distributions and yet contrast:

Monday/ Tuesday/ Wednesday/ Thursday/ Friday/ Saturday/ Sunday

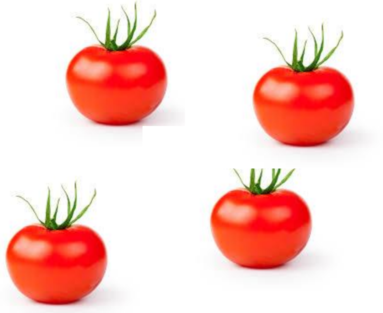
first/second/third/fourth/fifth/sixth/seventh/eighth/ninth/tenth

Manchester/Liverpool

Wide-ranging problem

- applies to any problem that uses distributional embeddings
  - inference north =/=>south
  - question answering *What is the third day of the week?*
  - image retrieval

five tomatoes=/=



etc

Capturing Discriminative Attributes

Given two related words, can we find what distinguishes them?

Semeval 2018 Task 10

With Alicia Krebs and Alessandro Lenci

- Given the words *apple* and *banana*, is *red* a discriminative attribute?

- 5K manually validated triples of the form <apple,banana,red>

https://aclanthology.org/S18-1117.pdf

Semeval 2018

With Alicia Krebs and Alessandro Lenci

- positive examples:

$word_1$	$word_2$	$attribute$
airplane	helicopter	wings
bagpipe	accordion	pipes
dolphin	seal	fins
gorilla	crocodile	bananas
oak	pine	leaves

- negative examples:

$word_1$	$word_2$	$attribute$
tractor	scooter	wheels
crow	owl	black
squirrel	leopard	fur

- upper bound: 90%
- leader: 75%
- cosine: 61%

Lessons

- Discriminative attributes is a hard problem for embedding models’16

- Big human/system gap

- Identifying semantic differences is difficult
- But maybe distributions still capture meaning distinctions?

- Shall we see this in text generation/prediction/language modeling?

Are large language models better?

- We'll take the example of GPT-2 and ChatGPT

- "Language Models are Unsupervised Multitask Learners"

(Radford et al. 2019)

Example (from Natural Questions dataset):

Who wrote the book the origin of species?

Generated answer: Charles Darwin

GPT-2

(source: [Write With Transformer (huggingface.co)]())

the third day of the week is

- Monday, the seventh day of the week is
- the feast of Saint Anthony, in the evening of
- the Feast of Epiphany, or the Feast of

the second day of the week is

- Thursday, the first day of the week is Friday
- called the Sunday, the
- called the Lord's day, and it was the

ChatGPT improves greatly:

What is the third day of the week?

The third day of the week is Wednesday. The names of the days of the week in English are Sunday, Monday, Tuesday, Wednesday, Thursday, Friday, and Saturday. The names of the days are often abbreviated to their first three letters, so Wednesday is commonly abbreviated to "Wed."

How about the discriminative attributes?

- ChatGPT:
  - doesn’t solve the problem zero shot
  - seems to contain all relevant information

Does having fur distinguish squirrels from leopards?

Yes, having fur is a key physical characteristic that distinguishes squirrels from leopards. Both squirrels and leopards are mammals, so they both have fur, but the type and appearance of their fur is different. Squirrels are small, agile rodents that belong to the family Sciuridae. They are characterized by their

GPT-2 example – where is Liverpool?

Liverpool is to the north of

England

London

Liverpool is to the south of

London, bu

the border

Newcastle i

Liverpool is to the east of

London, which is the city whe

England.

Paris, and has been for years

Liverpool is to the west of

London.

the River Mersey,

GPT-2: Manchester vs Liverpool

Are Manchester and Liverpool the same city?

Yes.

They are actually quite different cities,

I don't think so.

How about newer, bigger models? E.g. ChatGPT

What city is Manchester west of?

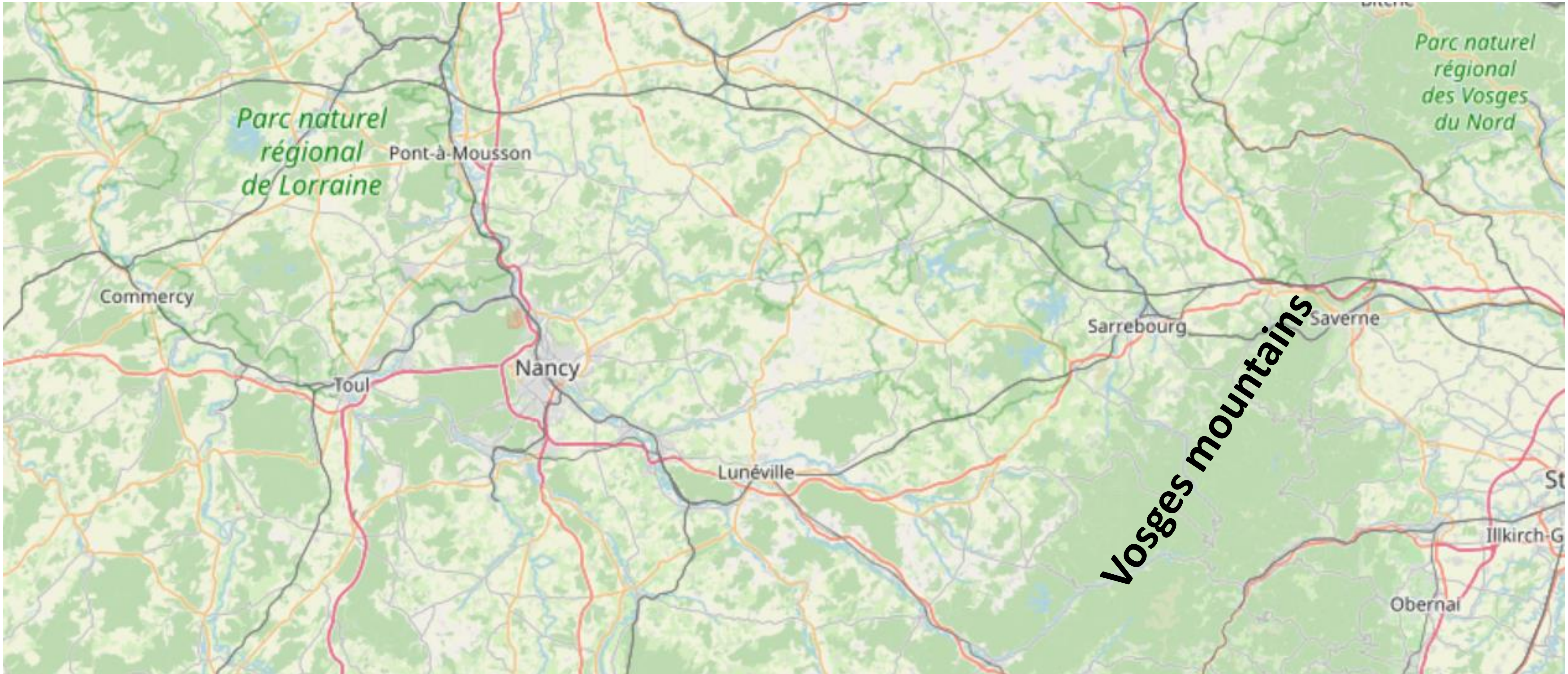
As I mentioned earlier, Manchester is located in the northwest of England. It is not west of any city in particular, as it is located in the northern part of the country. However, some cities that are located to the west of Manchester include Liverpool, Chester, and Preston.

ChatGPT: Correct facts from training data

What city is Manchester east of?

Again, Manchester is located in the northwest of England and is not east of any city in particular. However, some cities that are located to the east of Manchester include Leeds, Sheffield, and Nottingham. It's worth noting that Manchester is a large city and its eastern boundary is quite far from these cities, so they are not directly adjacent to Manchester.

En passant par la Lorraine



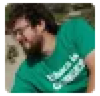
Is Nancy to the east of Toul?

Yes, Nancy is located to the east of Toul. Nancy is a city in the northeast of France, and it is the capital of the Lorraine region. Toul is a smaller town located in the same region, and it is situated to the west of Nancy. The distance between the two cities is about 30 miles, and they are connected by the A31 motorway.

Is Nancy to the east of the Vosges?

Yes, Nancy is located to the east of the Vosges mountains. The Vosges mountains are a range of low mountains located in the northeast of France, and they stretch from the Lorraine region in the west to the Alsace region in the east. Nancy is located in the Lorraine region, and it is situated on the eastern edge of the Vosges mountains. The city is about 30 miles from the nearest part of the mountain range, and it is connected to the Vosges by the N57 road.

ChatGPT: asking something very specific



Is Hatton Country World located to the east of Horton village?

I'm sorry, but I am not able to browse the internet or provide specific information about the location of Hatton Country World or Horton village. As a language model trained by OpenAI, my knowledge is based solely on the text that I have been trained on, and I do not have access to real-time information or the ability to browse the web.

ChatGPT

- Tells apart east from west and does some reasoning

- Memorizes lots of facts from training data

  Toul is a district administration and situated along the Moselle river and the Marne-Rhine canal. It is situated at a 15 minutes' drive West of Nancy, in the middle of the "Côtes de Toul" vineyards on the "Route du Vin et de la Mirabelle". - www.toul.fr - Presentation of the town

- Hallucinates/makes wild guesses/gives up when direct statements of facts are absent from training data

Probing the distributional hypothesis

All distributional models do this

$\Pr(t1 \mid c) > \Pr(t2 \mid c)$

Example:

c : She was in the ___.

t1: office

t2: Saturday

Do distributional models do this?

$\Pr(t1 \mid c) > \Pr(t2 \mid c)$

Example:

c : She was in the office last ___.

t1: Friday

t2: Saturday

Challenge: Can you know the word from the company it keeps?

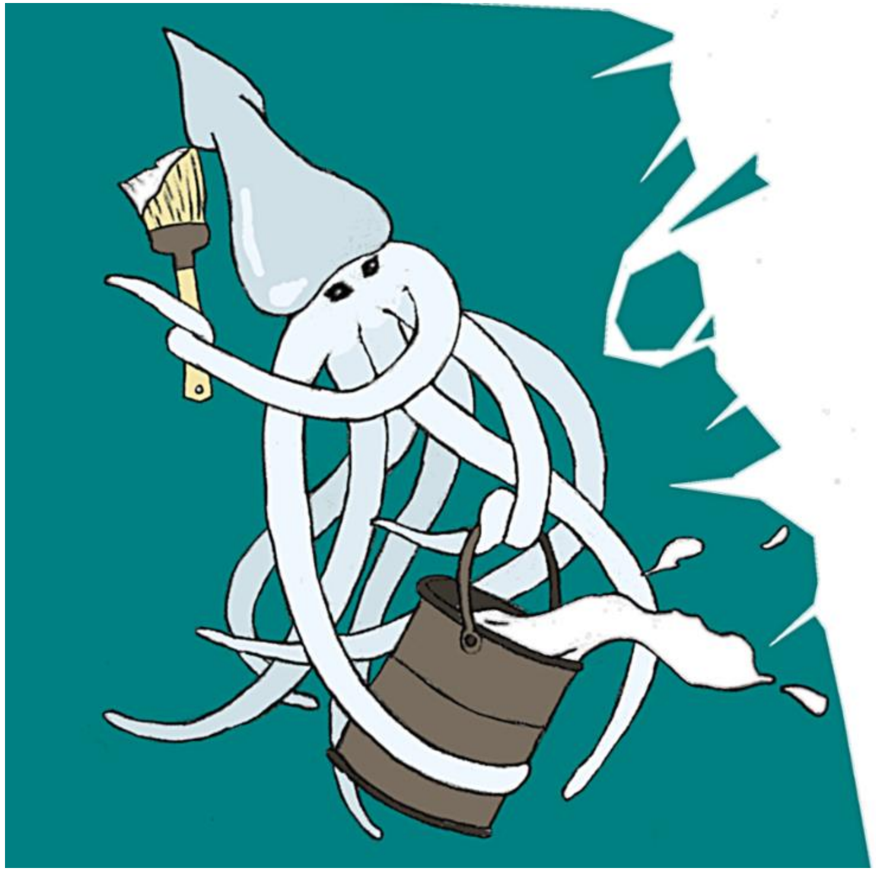
Given a word’s distribution, is it possible to identify the word/the lexical meaning?

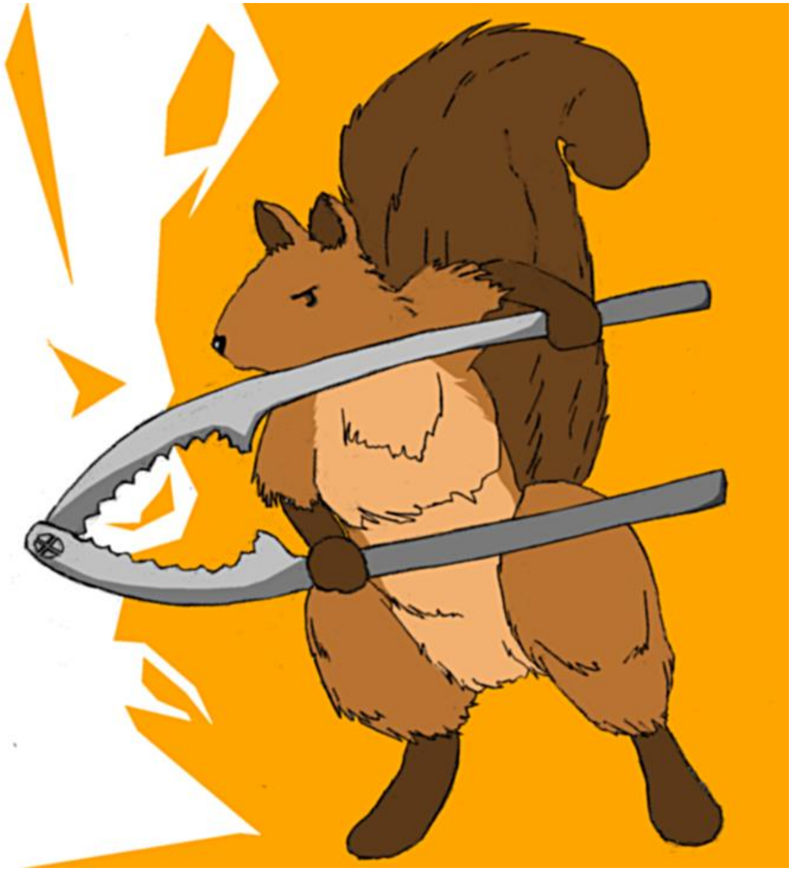
- (SAMPLE of contexts of Saturday) > `Saturday’?

BlankCrack experiment

https://blankcrack.atilf.fr/game/
https://arxiv.org/abs/2108.07708

- with Timothee Mickus and Mathieu Constant



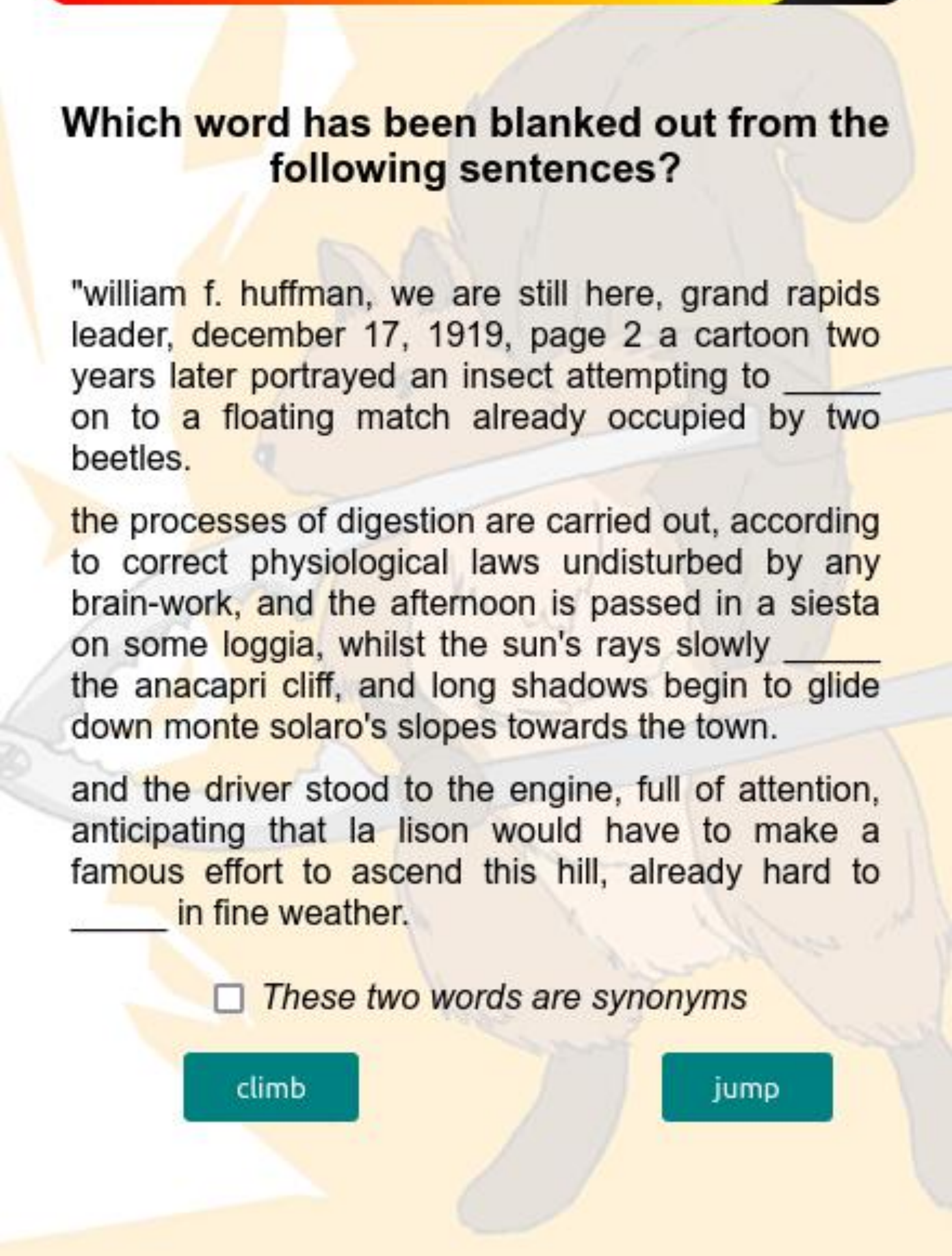


The BlankCrack experiment

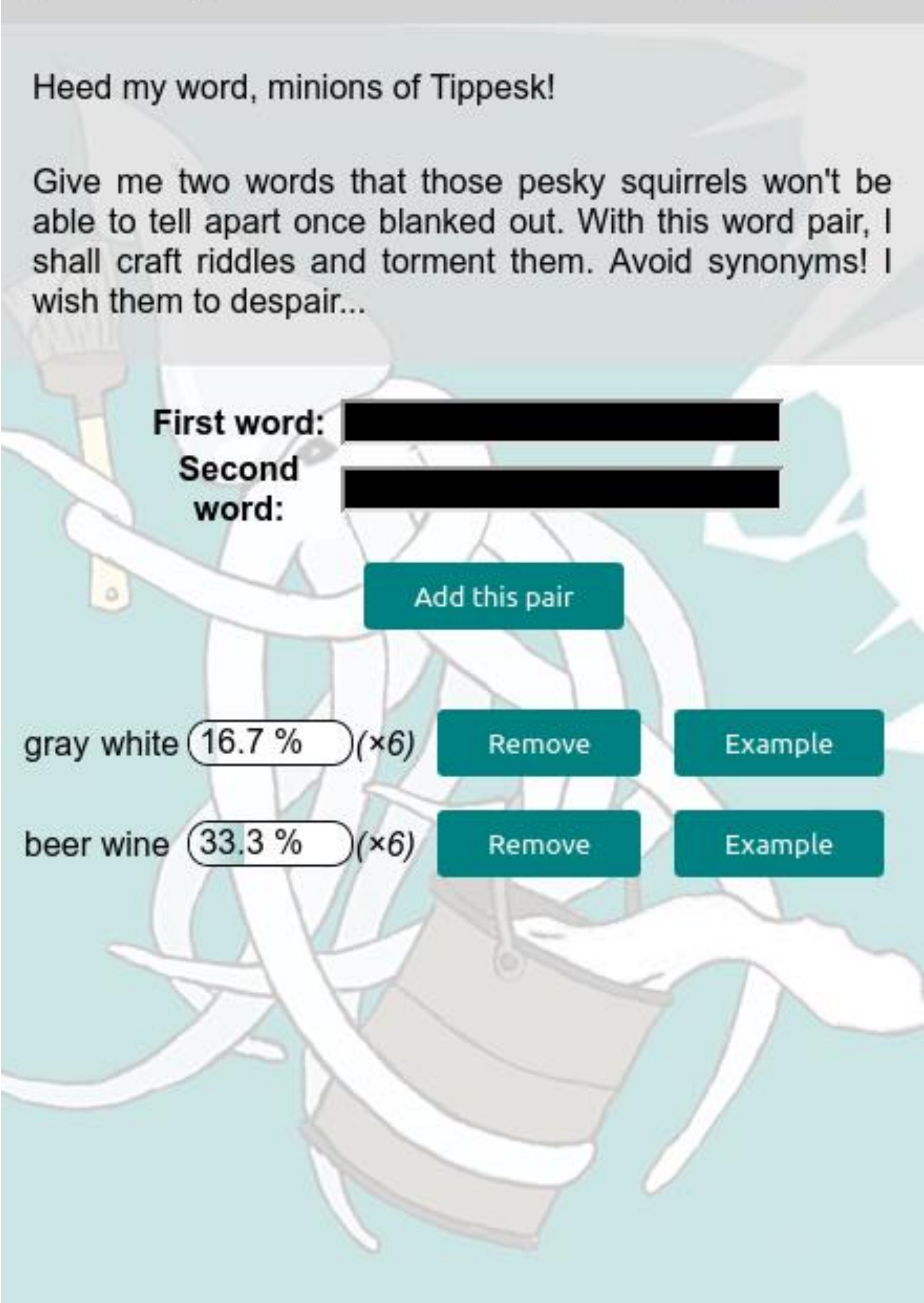
- Goal: identify meaning contrasts that evade distribution, if any
- How: collect human judgments
- Method: gamify!

https://blankcrack.atilf.fr/game/

The game interface: resolving the word's identity from its distribution



Players can propose their own word pairs



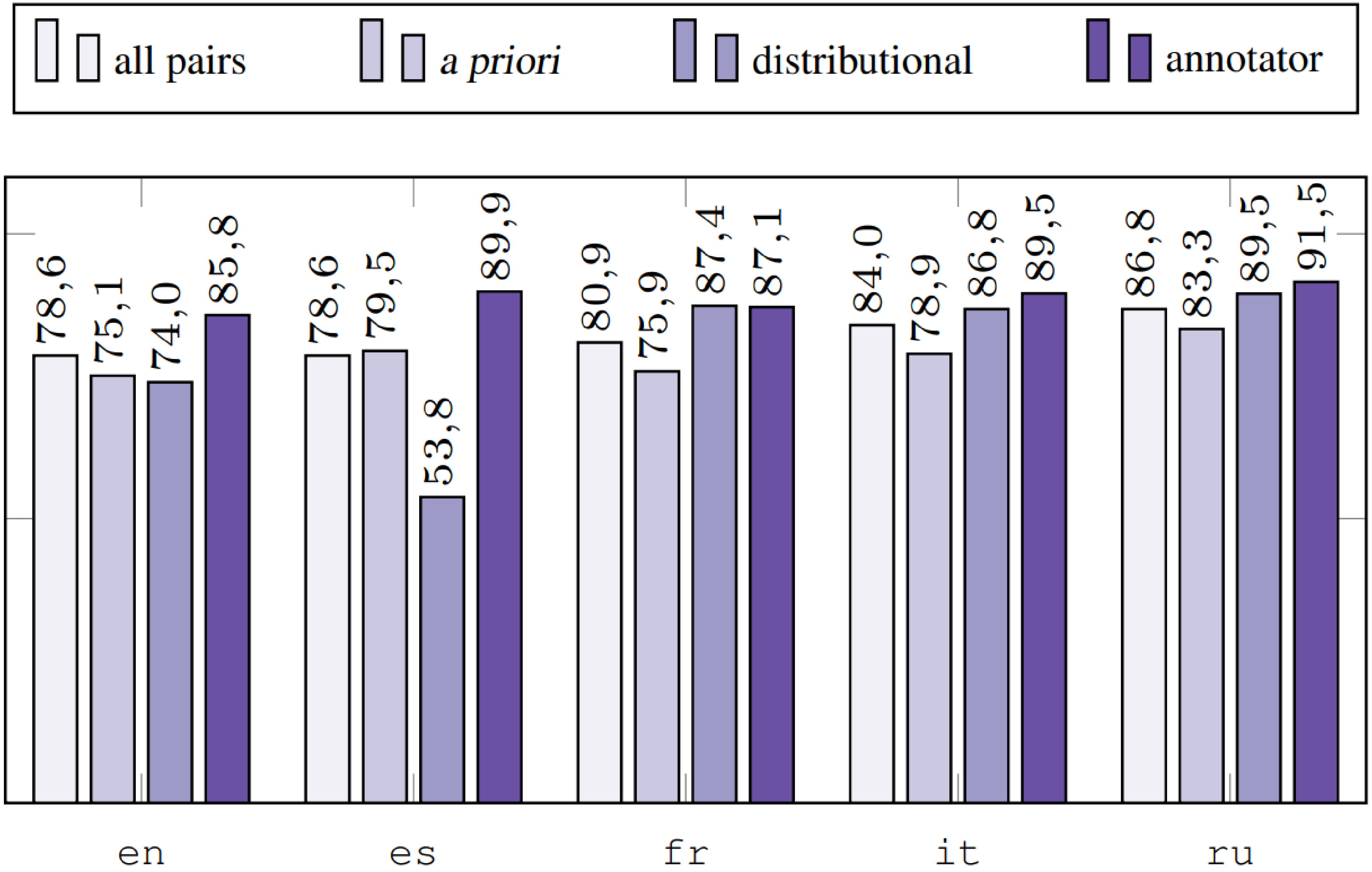
Word pair types

- participant suggested pairs
- distributional neighbors
- manual (a priori)
  - months: October vs July
  - numbers: three vs five
  - colors: black vs grey
  - days of the week Wednesday vs Saturday

Number of annotations collected

	`en`	`es`	`fr`	`it`	`ru`
$k=1$	329	110	540	161	113
$k=3$	58	90	136	73	90
$k=5$	2223	2044	3719	816	3991
Total	2610	2244	4395	1050	4194

Success rates (humans):



Models

- Baselines
  - unigram
  - bigram

- Embedding models (pretrained)
  - word2vec
  - BERT (BERT/BETO/UmBERTo/CamemBERT/ruRoberta)

Success rates (models)

	en	es	fr	it	ru
human	83.1	86.9	83.8	89.1	87.8
1-gram	51.9	56.2	53.4	50.8	57.2
2-gram	60.4	71.2	66.0	70.7	60.1
BERTs	75.8	71.6	74.1	76.1	74.4
W2Vs	75.5	77.1	75.5	74.8	72.5

Examples of indistinguishable pairs

Most word contrasts are distributional (>80% in our biased sample!)

Some words are special

distributional signal is weak

require extra **knowledge of properties of referents**

- hyena & jackal, baseball & basketball (English)
- aquarelle & gouache (French)
- cilantro & cebollino (Spanish)

Training on more text data can provide this knowledge.

Beyond distributional limits

- For some aspects of meaning, visual input clearly helpful
  - e.g. spatial relations
- Rare cases where language-vision correspondence is crucial
  - e.g. Winoground dataset

  (Thrush et al. 2022)

(a) some plants surrounding a lightbulb

(b) a lightbulb surrounding some plants

- Much of the time, visual information can also be obtained from text (e.g. color of objects)

Beyond distributional limits: Geo-Aware Image Caption Generation

Sofia Nikiforova, Tejaswini Deoskar, Denis Paperno and Yoad Winter

Dissertation work of S.Nikiforova within the ERC AdG ROCKY project (PI Y.Winter)

https://aclanthology.org/2020.coling-main.280/

Example

Figure 1: An example image.

Ground Truth: A path through Pitshanger Park, near Ealing in the west London suburbs
Automatically generated: a park bench sitting in the middle of a park

EXAMPLE

Ground truth caption:

Grand Union Canal locks near Hatton Country World taken on a wet day

EXAMPLE

Standard captioning system

(Xu et al. 2015):

the bridge carries the over the canal just west of horton village

EXAMPLE

Standard captioning system

(Xu et al. 2015):

the bridge carries the over the canal just west of horton village

Canal Road, Hatton, Warwick, Hatton Park, W

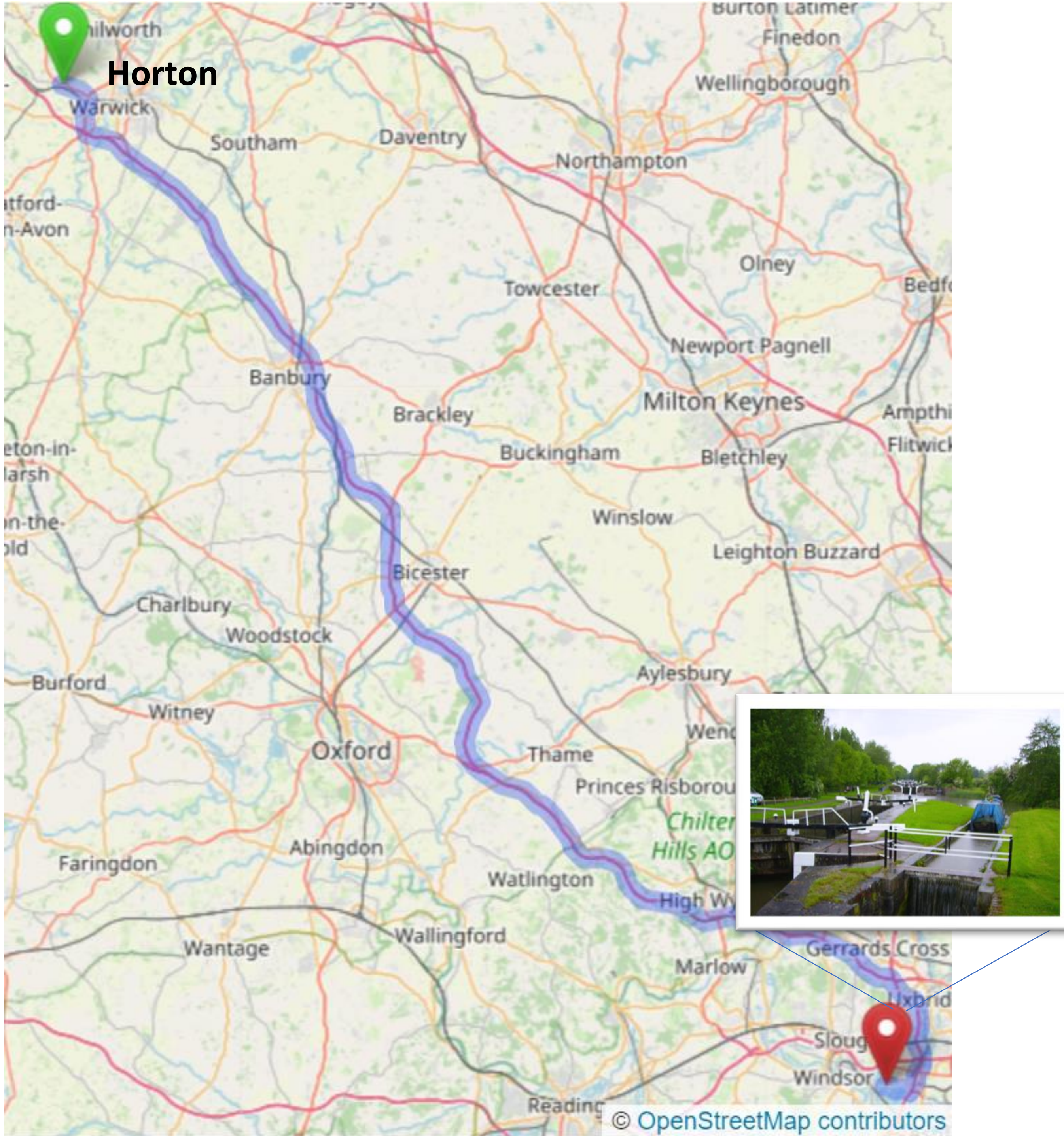
Horton, Royal Borough of Windsor and Maid

Car (OSRM) Go

Reverse Directions

Directions

Distance: 141km. Time: 1:38.



Proposal

- **General vocabulary** has distributional vectors
- **Special items** get vectors by embedding their properties (from KB)

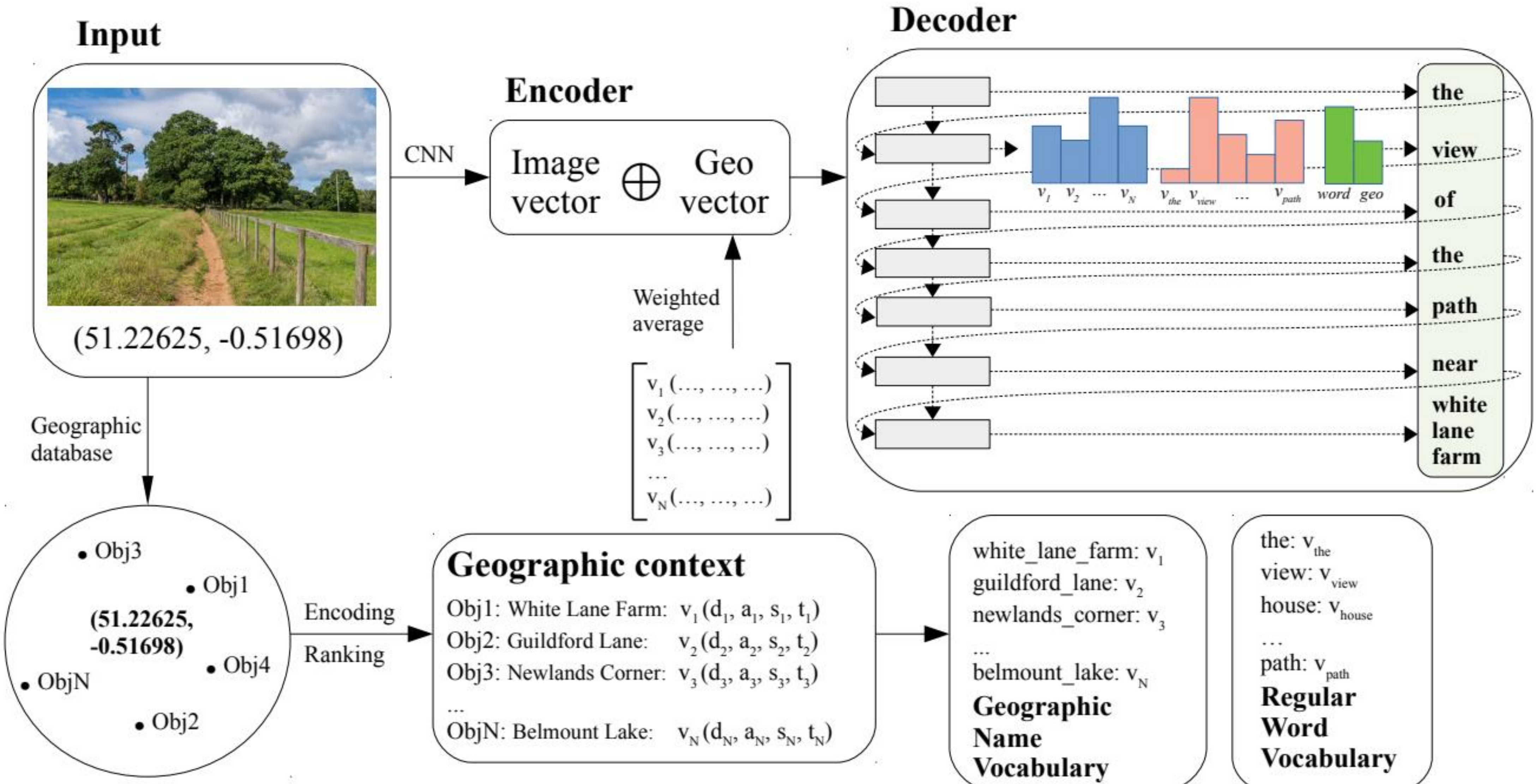
Name	Type	Size, km^2	Latitude	Longitude
Cambourne	town	7.264	52.219984	-0.070078

- For example, embedding geographic entities:

$$\textsc{GeoEmb}(o_i) = d_i\vec{w_d} + a_i\vec{w_a} + s_i\vec{w_s} + E_t(t_i)$$

  - d – distance
  - a – azimuth
  - s – size
  - E_t – type (village, road, river etc.)

more complete architecture



EXAMPLE

Ground truth caption:

Grand Union Canal locks near Hatton Country World taken on a wet day

Standard (Xu et al. 2015):
the bridge carries the over the canal just west of horton village

Our system:

the view of the lock on the grand union canal near hatton

Quantitative results

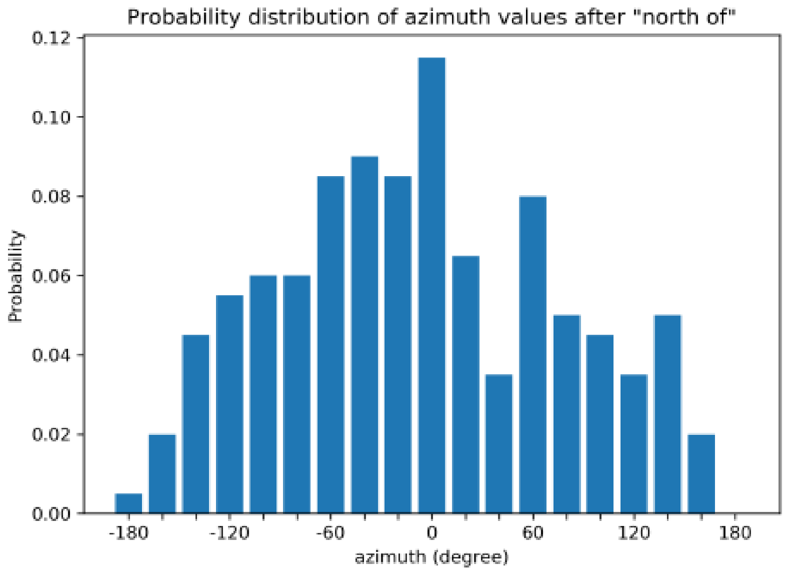
- Dataset: GeoRic, 29K images from geograph.org.uk
- average of 2 geographic entities per caption

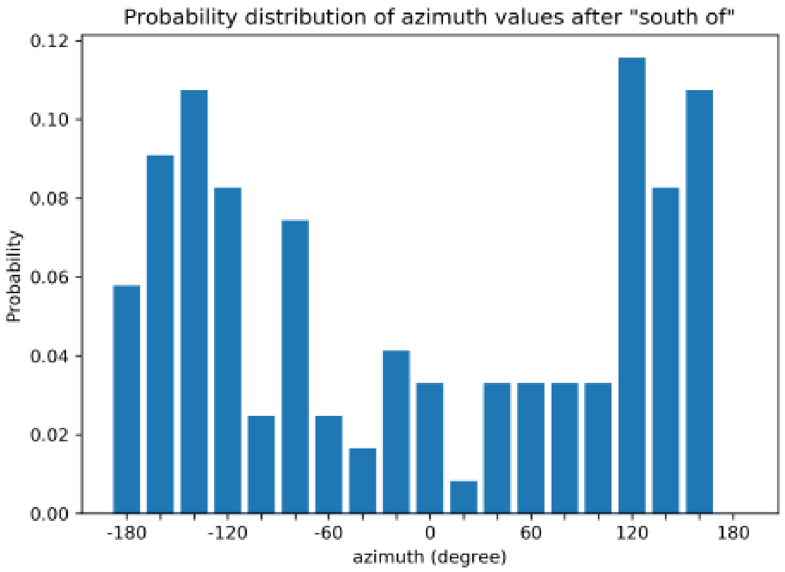
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	METEOR	CIDEr
Standard	13.38	2.82	0.64	0.33	15.79	5.55	7.38
Geo-aware	**18.12**	**8.42**	**3.42**	**1.46**	**22.61**	**10.35**	**70.53**

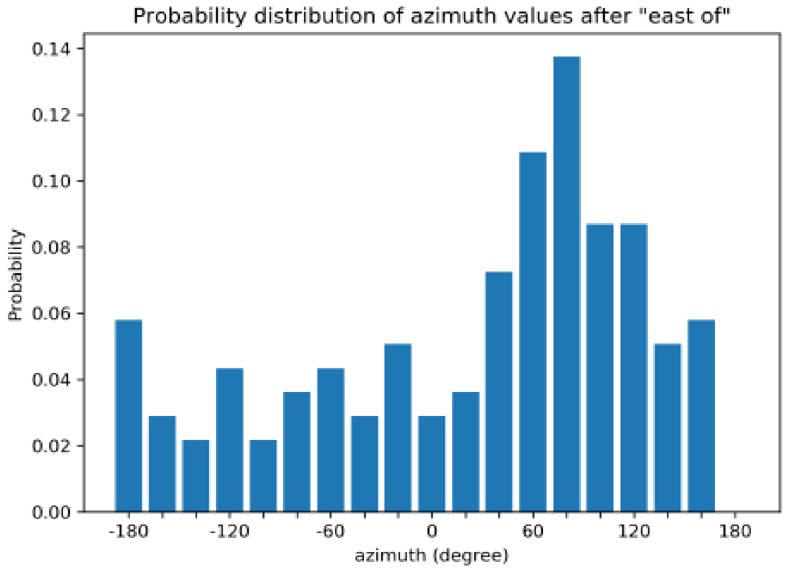
East or West

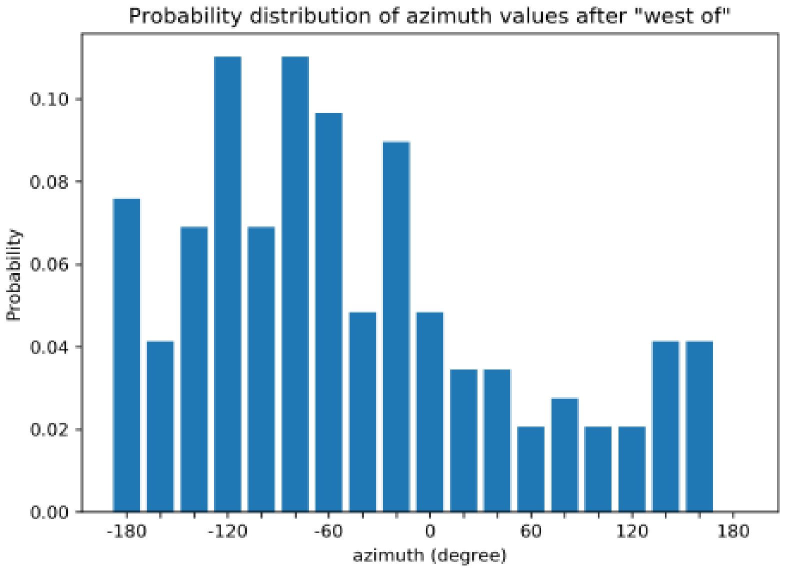
- Knowledge based embeddings of entities allow the system to learn directions

from Nikiforova,S. Dissertation manuscript. 2022









Follow-up: generalization to other knowledge

Ground truth: Kelso Bridge. Below the confluence of the Rivers Tweed and Teviot. John Rennie engineered the bridge, which was built between 1800 and 1803.

Standard (Mokady et al., 2021): A river with a bridge and a train on it.

No-knowledge: the *river dee*. *farndon bridge* was opened in *1339* by *monks* from *farndon bridge*.

Knowledge-from-image: *chertsey bridge*. *chertsey bridge* dates from *1785*.

Knowledge-from-metadata (ours): **kelso bridge**. the bridge over the **river tweed** was built in **1800**, and was designed by **john rennie the elder**.

Sofia Nikiforova, Tejaswini Deoskar, Denis Paperno, Yoad Winter

Generating image captions with external encyclopedic knowledge

https://arxiv.org/abs/2210.04806

Conclusion

- Semantics is >>80 distributional
- The rest should be grounded in reference and knowledge
- Geo-embeddings: successful non-distributional embeddings

Contact: denis.paperno@gmail.com

d.paperno@uu.nl

Thanks to all my collaborators:

Mathieu Constant

Tejaswini Deoskar

Alessandro Lenci

Alicia Krebs

Timothee Mickus

Sofia Nikiforova

Yoad Winter

Thank you!