To the Limits of Distributional Semantics and Beyond

Denis Paperno

7.7.2022

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 models

Very different architectures:

- matrix decomposition (Landauer and Dumais, 1997)
- log-linear classifiers (Mikolov et al., 2013)
- recurrent neural networks (Peters et al., 2018)
- transformers (Devlin et al., 2018)

Very different objectives:

- Multinomial or binomial classification (Mikolov et al., 2013)
- Language modeling (Bengio et al., 2003; Peters et al., 2018; Radford, 2018)
- Sequence denoising (Devlin et al., 2018; Lewis et al., 2020)

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

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
- Big human/system gap

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

- Shall we see this in text generation/prediction?

Example from generation

From definition generation work with T.Mickus and M.Constant

generating a definition for “Wednesday”

- target: “the fourth day of the week”
  (Source: merriam-webster.com)
- metrics will favor generated “the tenth day of the week”
- over the attested “The fourth day of the week in many religious traditions, and the third day of the week in systems using the ISO 8601 norm; it follows Tuesday and precedes Thursday.” (Source: en.wiktionary.org)

Are large language models better?

- We'll take the example of GPT-2

- "Language Models are Unsupervised Multitask Learners"

(Radford et al. 2019)

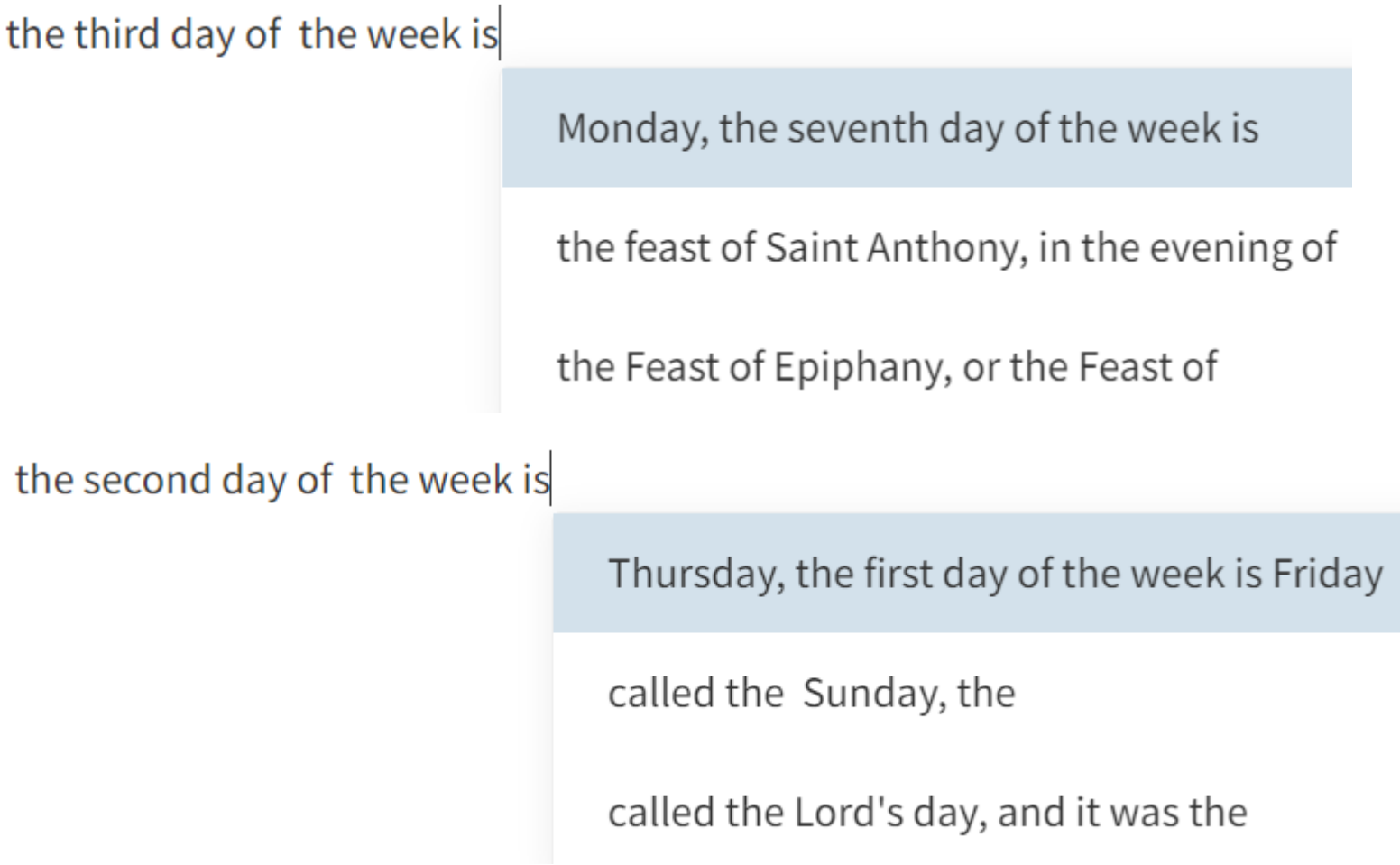
Example (from Natural Questions dataset):

Who wrote the book the origin of species?

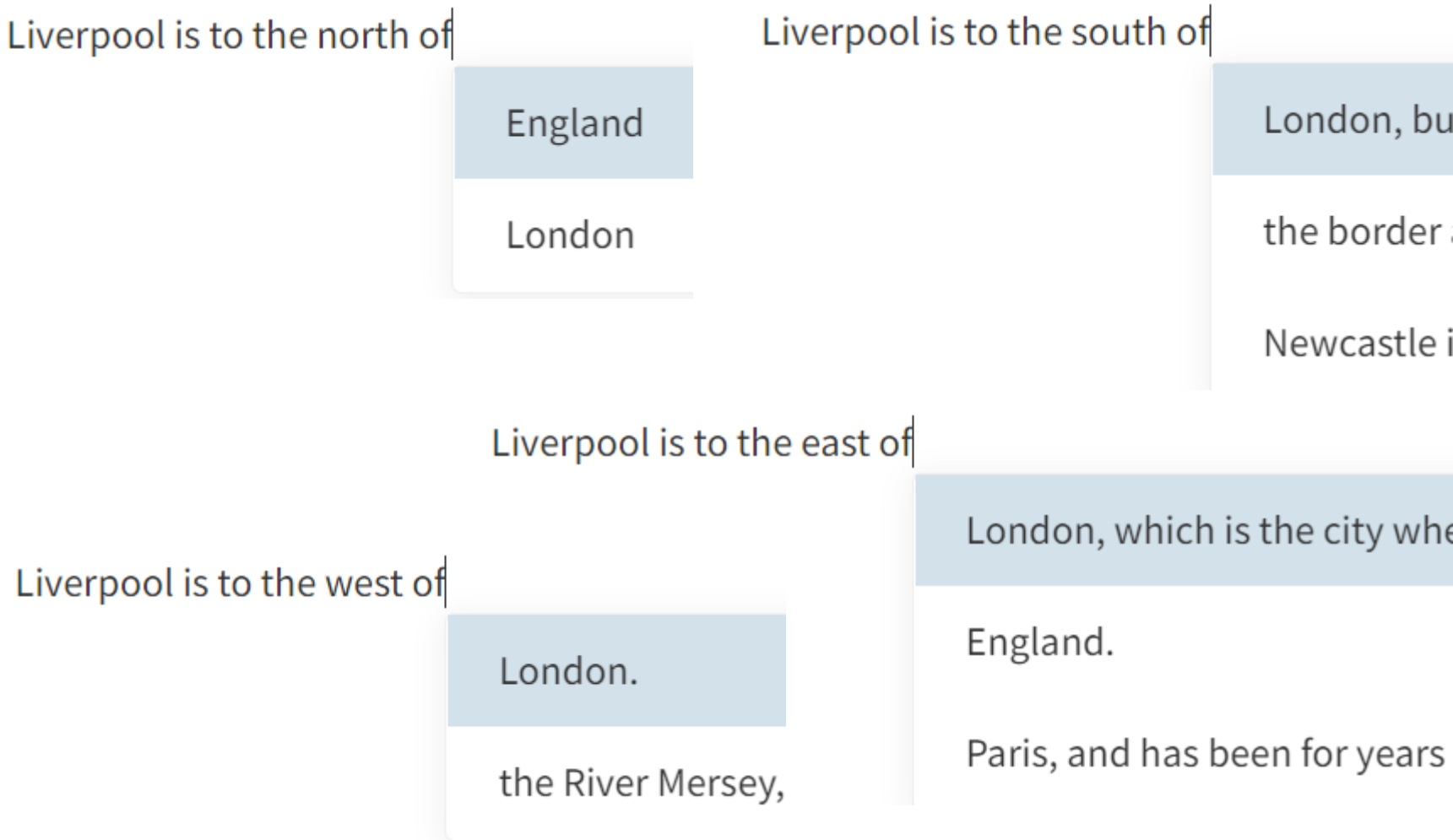
Generated answer: Charles Darwin

GPT-2

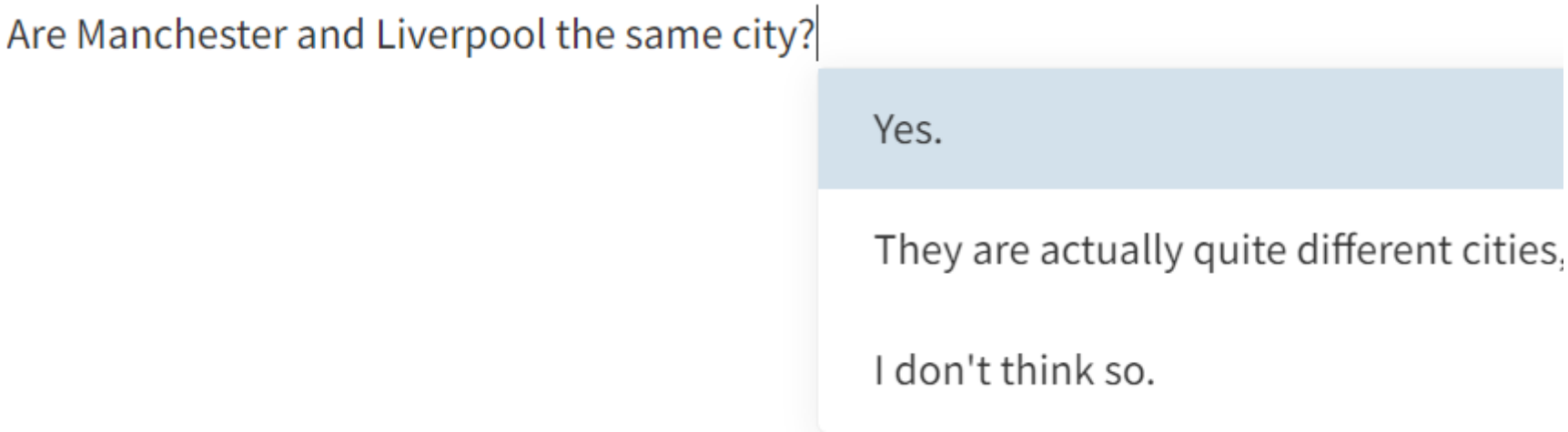
(source: Write With Transformer (huggingface.co))



GPT-2 example – where is Liverpool?



GPT-2: Manchester vs Liverpool



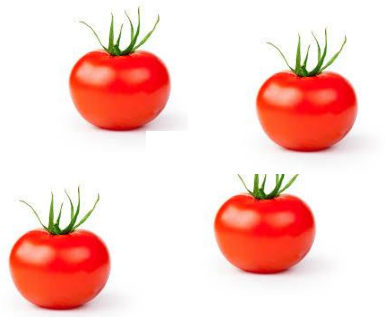
Wide-ranging problem

- examples from generation
- but applies to any problem that uses distributional embeddings
  - inference
  - question answering
  - image retrieval

north =/=>south

What is the third day of the week?

five tomatoes=/=



etc

Probing the distributional hypothesis

All distributional models do this

Pr($t1$ |c) > Pr($t2$ |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?

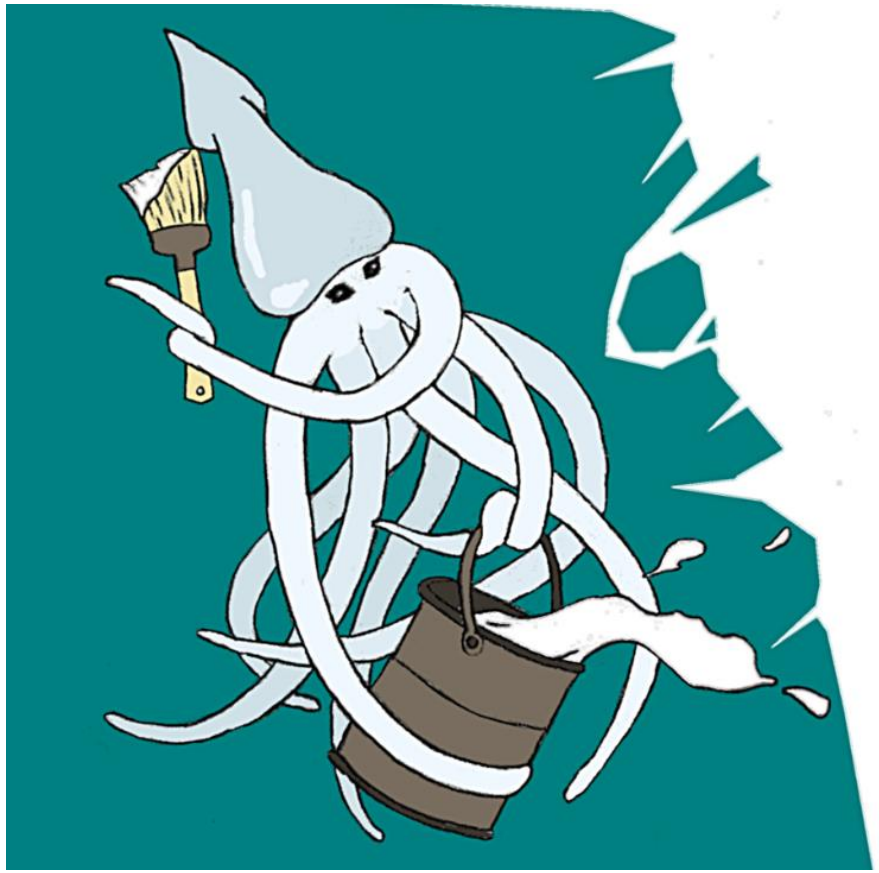
Given a word's distribution, is it possible to identify the word/the lexical meaning?

- (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

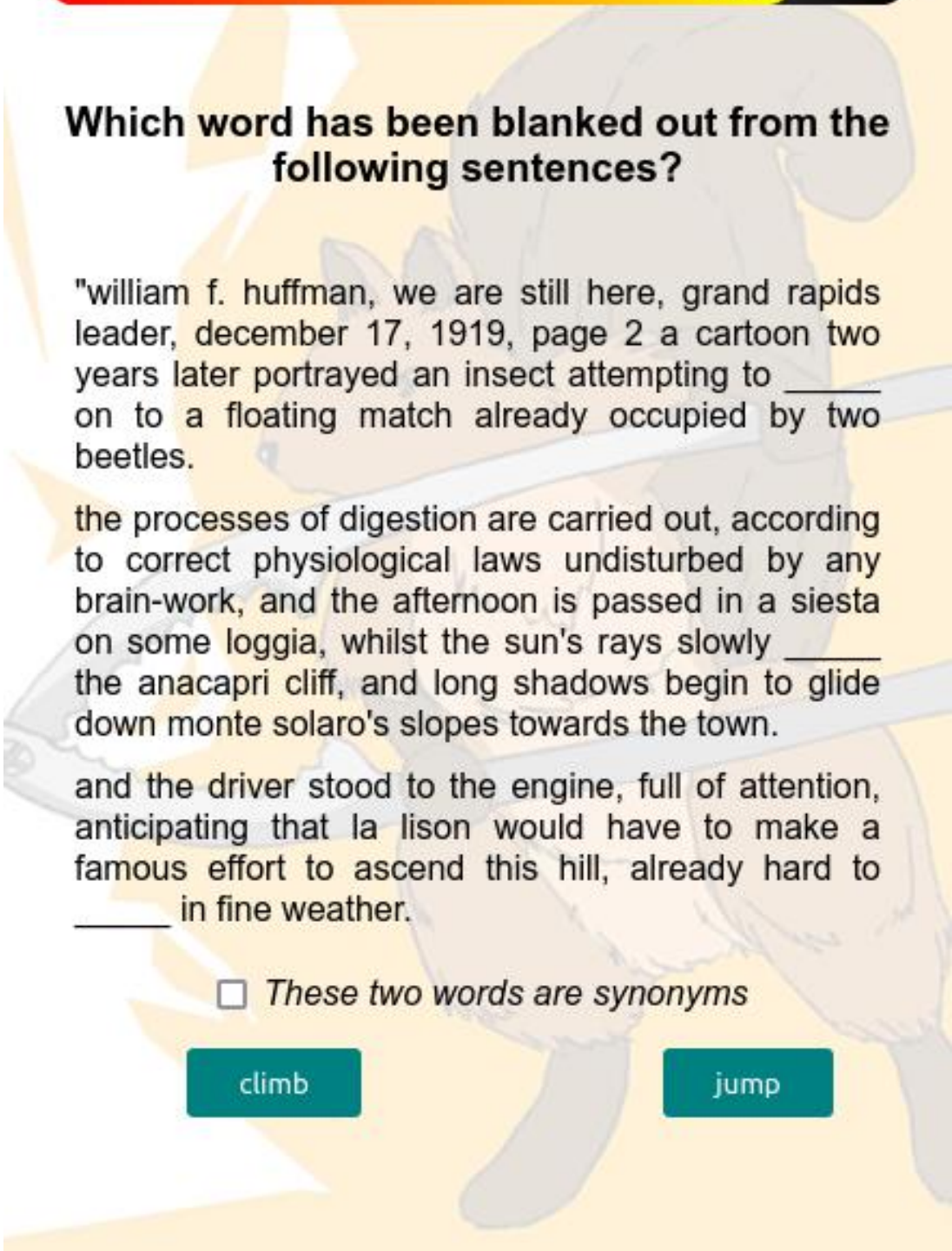


The BlankCrack experiment

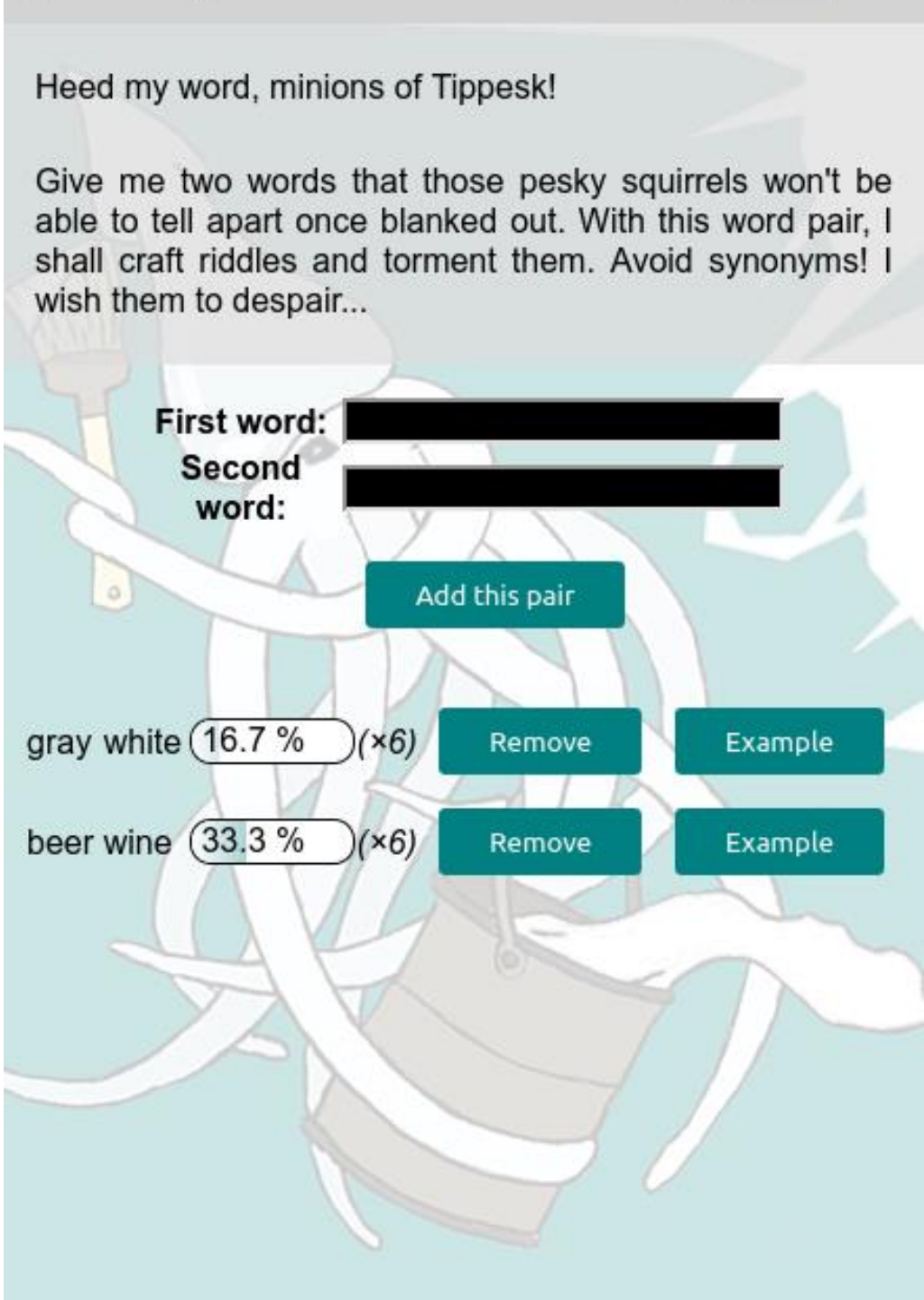
- Goal: identify meaning contrasts that evade distribution, if any
- How: collect human judgments
- Method: gamify!

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

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



Players can propose their own word pairs



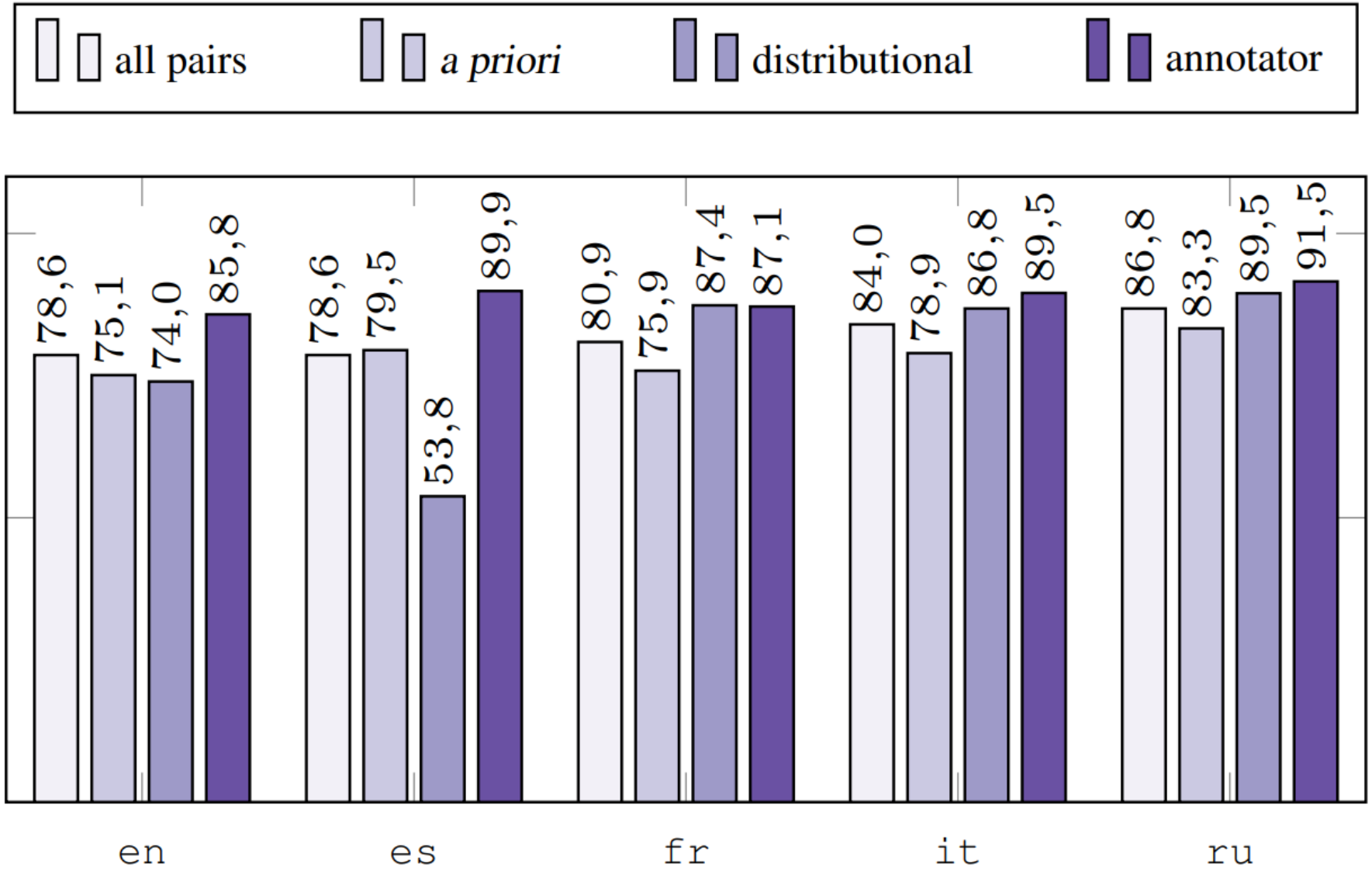
Word pair types

- 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 and require extra **knowledge**

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

Beyond distributional limits

with Sofia Nikiforova, Tejaswini Deoskar and Yoad 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):

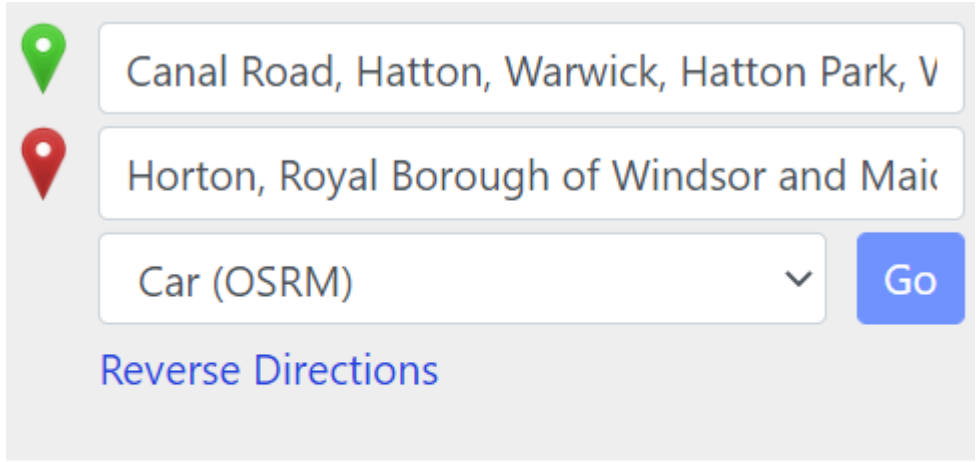
the bridge carries the over the canal just west of horton village

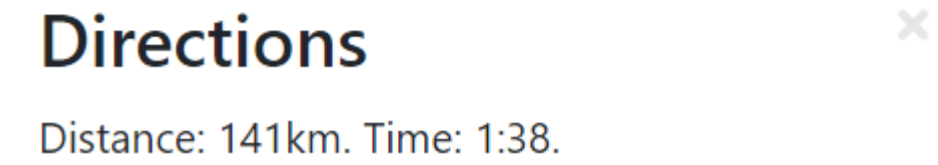
EXAMPLE

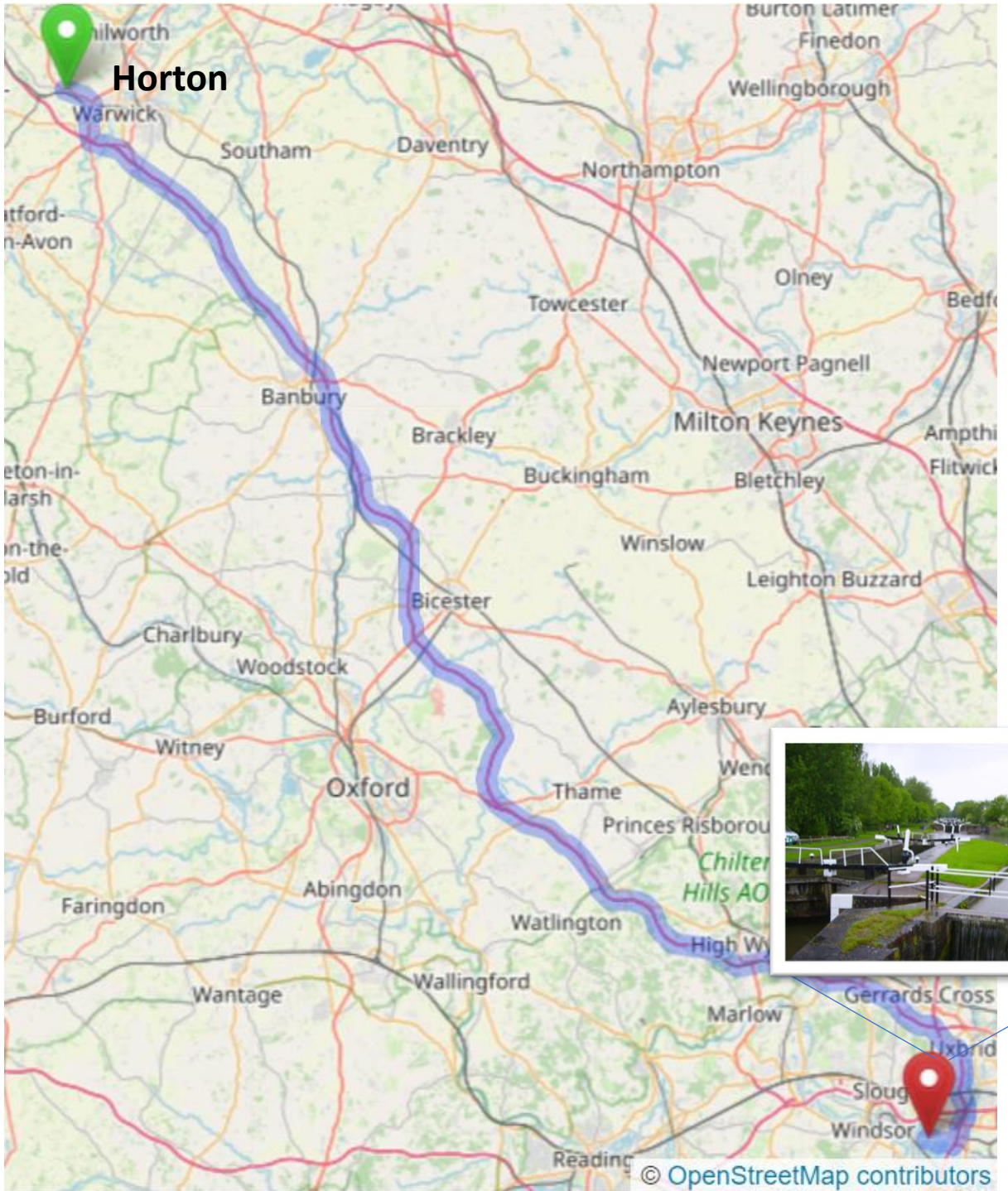
Standard captioning system

(Xu et al. 2015):

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







Proposal

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

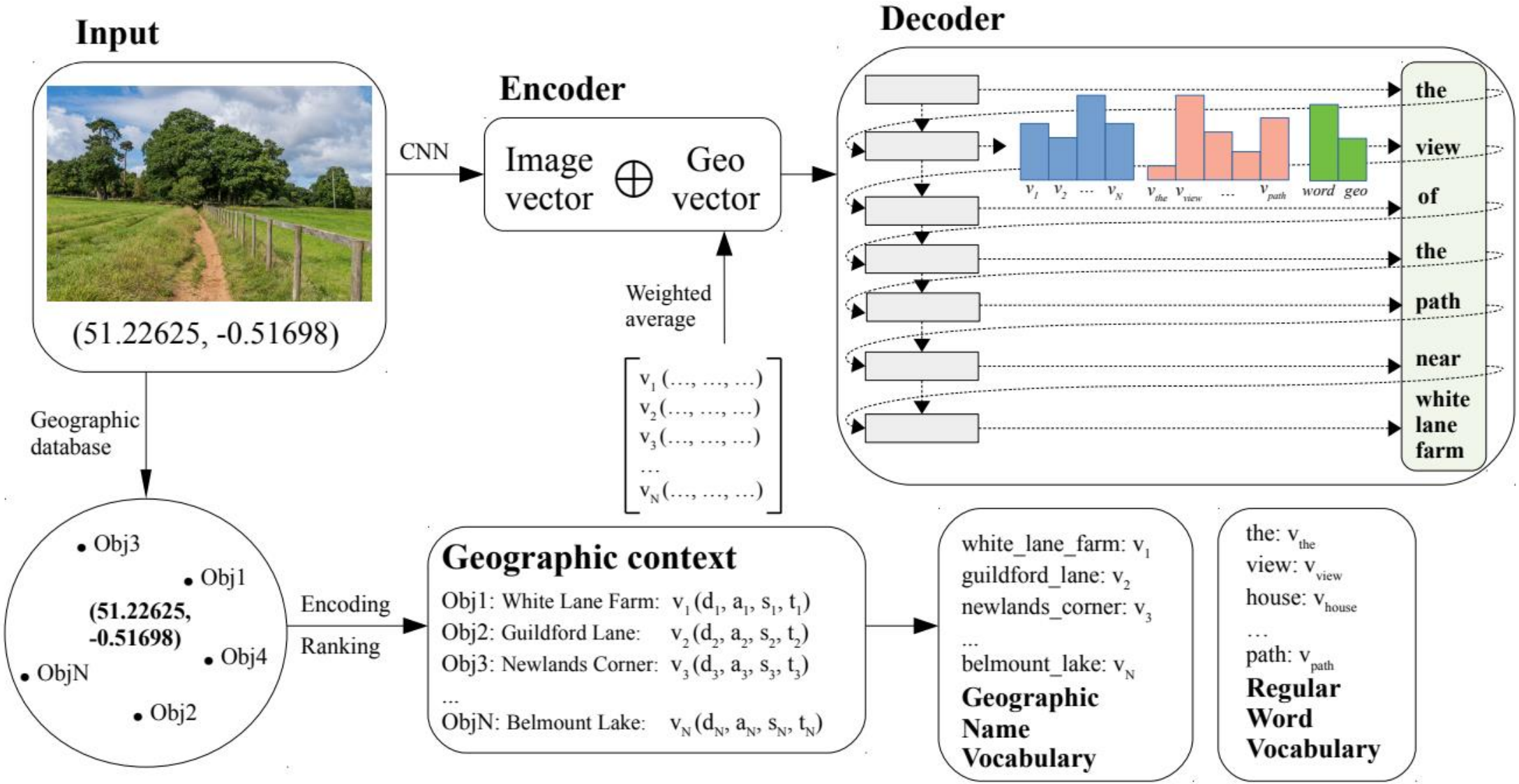
Name	Type	Size, km^2	Latitude	Longitude
Cambourne	town	7.264	52.219984	-0.070078

- 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**

Conclusion

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

- Open problem: extension to other types via knowledge bases

  e.g. BDPedia:

  - hyena: `Category:Carnivorans_of_Africa`
  - jackal: `Category:Carnivorans_of_Asia`

Contact: denis.paperno@gmail.com

Thank you!

Thanks to all my collaborators:

Mathieu Constant

Tejaswini Deoskar

Alessandro Lenci

Alicia Krebs

Timothee Mickus

Sofia Nikiforova

Yoad Winter