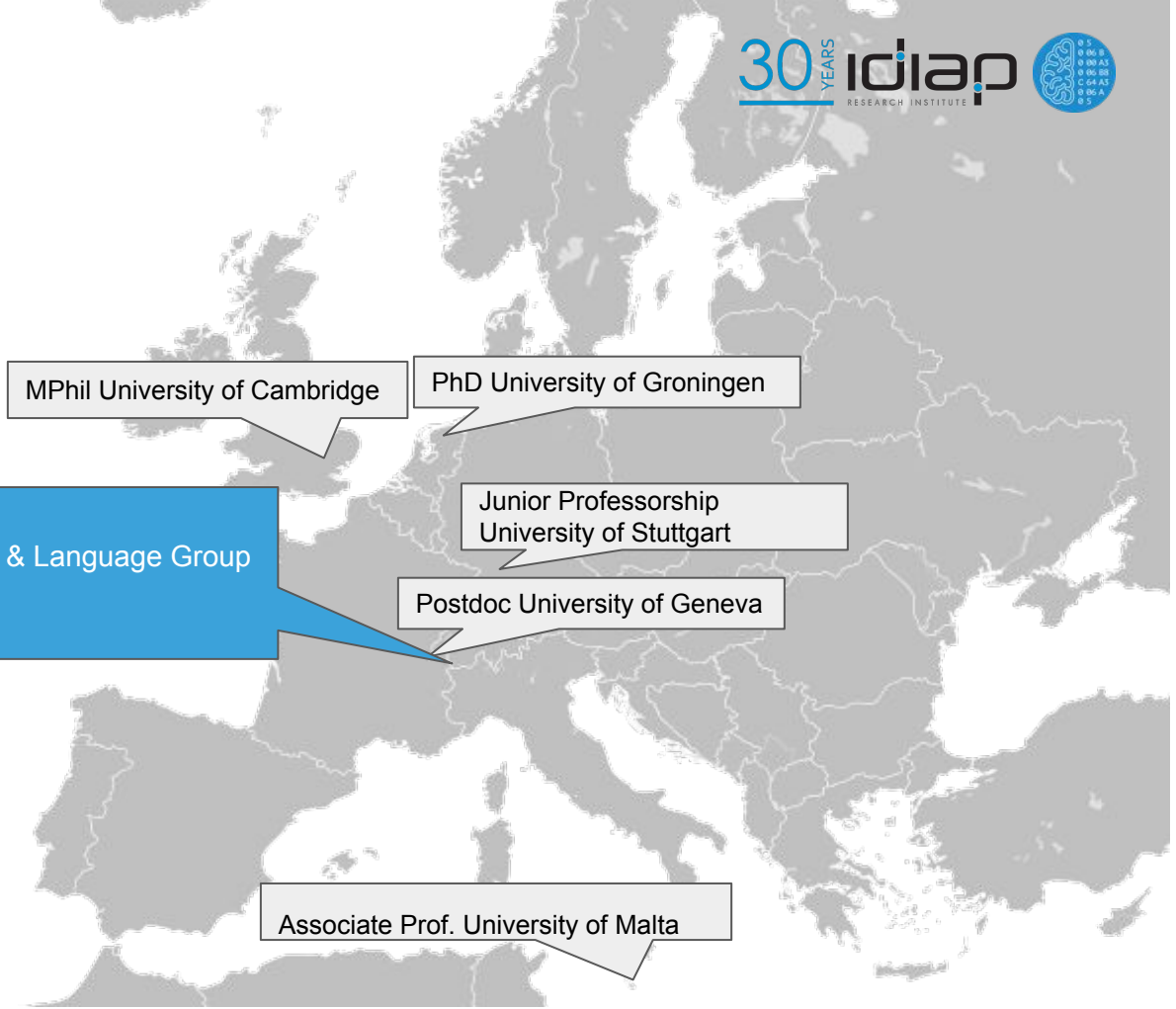
Crossing boundaries: from cross-lingual learning to creative thought

Lonneke van der Plas, Idiap

Joint work with Inga Lang , Marc Tanti, Claudia Borg, Albert Gatt, and Prajit Dhar

Helsinki 17.3.2022

Who am I?



Currently leading the Computation, Cognition & Language Group at Idiap in Martigny

Artificial Intelligence for society

- Independent not-for-profit Research Foundation, created in 1991
- 14 research groups covering a broad range of AI research areas
- A dedicated R&D engineers team bridging the gap between academia and industry
- 4 technology transfer highlights
  - Torch Facebook PyTorch
  - KeyLemon + Recapp Swisscom TV speech recognition
  - Swiss Biometrics Center FIDO certification (only 7 labs in the world)
  - Master in Artificial Intelligence a business integrated university training program

Expertise

Signal Processing

Computer Vision

Robotics

Machine Learning

Speech & Language

Human Computer Inter.

Privacy & Security

Data Science

Data types

Text

Speech and Audio

Images

Video

...

idiap RESEARCH INSTITUTE

+150 employees, +65 research projects and +120 publications per year

Application domains

Health and Life Sciences

Energy

Security

Manufacturing and Industry 4.0

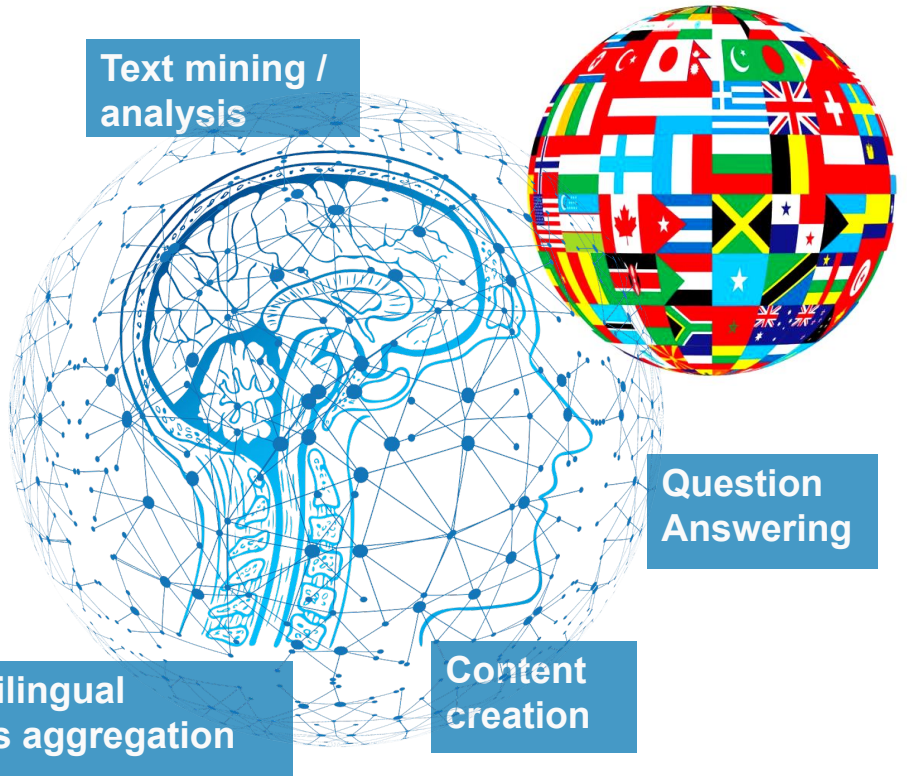
Media and Entertainment

Devices

Computation, Cognition & Language Group

Boundaries of current AI system with respect to **language**:

- Cross-lingual transfer for language technology
- Modelling **human cognitive abilities** that are underexposed, such as **creativity**



[Image adapted from Gerd Altmann from pixabay.com]

Limitations to current AI

- Recent trend has been to feed more and more data to learning methods
- This has led to impressive results in several tasks
- Also, awareness of limitations of these systems
- They are brittle, data-hungry, task-specific/narrow, unable to generalise beyond the training distribution, and not learning in a flexible way as humans do, opaque
- All-in-all they lack many aspects of human intelligence

Aim of my research

- Investigate methods to address previously mentioned shortcomings
- Crossing boundaries: across languages, from one tasks to the other, across domains, and across time spans, in low-resource scenarios
- Models inspired by human cognition

Cross-lingual transfer and computational creativity

Impact of fine-tuning on language-specificity of mBERT

Generating novel concepts

mBERT and the effect of fine-tuning

(work in collaboration with Marc Tanti, Claudia Borg and Albert Gatt)

BERT is a large pre-trained transformer model

It is trained on masked language modelling and next-sentence prediction

It captures the semantic similarity between words in context

Model is task-agnostic and can be used for several tasks in fine-tuning set up

mBERT is the multilingual counterpart of BERT (pre-trained from monolingual corpora in 100+ languages)

Language-specificity in mBERT

Recent work has suggested that mBERT has two components:

- a **language-specific** and
- a **language-neutral** one

The language-neutral component could explain why mBERT works surprisingly well when used for **zero-shot cross-lingual transfer**

Where a model is fine-tuned on monolingual annotated data and tested on data from another language, while both are included in mBERT

The effect of fine-tuning

Why does zero-shot transfer work?

We would need representations that cut across linguistic distinctions for this to work, right?

But mBERT is trained on texts in different languages without any cross-lingual objective

What happens exactly to mBERT's representations after fine-tuning?

Experiments

Two tasks:

- NLI (natural language inference)
- UDPOS (part-of-speech tagging)

> Fine-tune on English labelled data

- Check performance on task in non-English languages
- Check language-specificity of fine-tuned mBERT using:
  - t-SNE visualisations
  - language identification task (on target data and on separate test set)

Results for XNLI

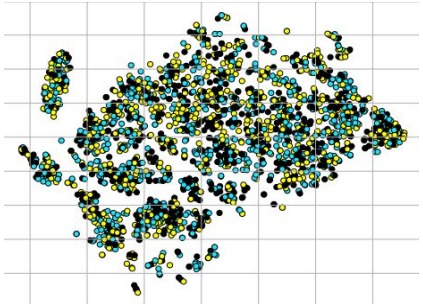


Fig 1: XNLI labels with initial mBERT
F1 XNLI labels: 29.7%

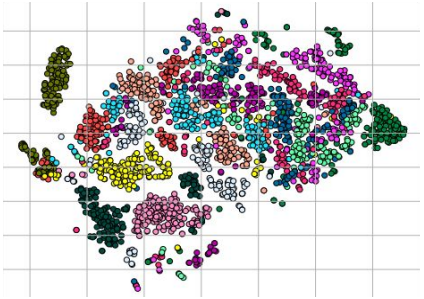


Fig 2: XNLI languages with initial mBERT
F1 XNLI langs: 49.8%

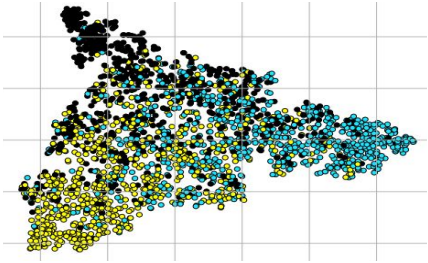


Fig 3: XNLI labels with fine-tuned mBERT
F1 XNLI labels: 66.3%

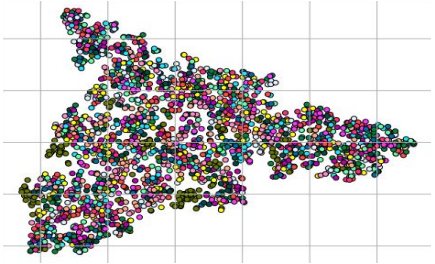


Fig 4: XNLI languages with fine-tuned mBERT
F1 XNLI langs: 39.2%

The effect of fine-tuning on two different tasks

	UDPOS		XNLI	
	Init.	Fine-T.	Init.	Fine-T.
Target task	51.2	59.6	29.7	66.3
Lang. ID (Target)	78.3	0.3	49.8	39.2
Lang. ID (Wiki)	59.3	0.5	97.0	97.2

Table 1: Macro F1 scores (%) for target tasks (UDPOS and XNLI) and language identification before (Init.) and after fine-tuning (Fine-T.). Note that 'Lang. ID (Target)' refers to language classification on the target data set.

The effect of fine-tuning on two different tasks

	UDPOS		XNLI	
	Init.	Fine-T.	Init.	Fine-T.
Target task	51.2	59.6	29.7	66.3
Lang. ID (Target)	78.3	0.3	49.8	39.2
Lang. ID (Wiki)	59.3	0.5	97.0	97.2

What we learn:

- XNLI gains most from fine-tuning
- UDPOS only gains some
- Fine-tuning for UDPOS leads to mBERT losing more of its language specificity

What that implies:

- finetuning requires mBERT's finite representational capacity to be dedicated to the task, at the expense of accurately distinguishing between languages

How about the differences between the tasks?

How about the differences between the tasks?

	UDPOS		XNLI	
	Init.	Fine-T.	Init.	Fine-T.
Target task	51.2	59.6	29.7	66.3
Lang. ID (Target)	78.3	0.3	49.8	39.2
Lang. ID (Wiki)	59.3	0.5	97.0	97.2

Why does training on a morpho-syntactic, token-level task lead to large decreases in language specificity?

Why does fine-tuning lead to more steep increases in performance for XNLI and less so for UDPOS?

Lauscher et al. (2020) show that POS tagging and dependency parsing are impacted by structural language similarity.

NLI is known to be susceptible to ‘shortcut learning’ (D’Amour et al.)

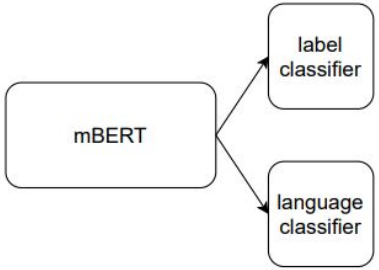
The data sets differ in their homogeneity and granularity

Discussion

If zero-shot transfer works because of representations that cut across linguistic distinctions,

can we change the learning architecture in such a way that it focuses more on language-independent features to benefit cross-lingual transfer?

Language-unlearning with mBERT I: gradient reversal



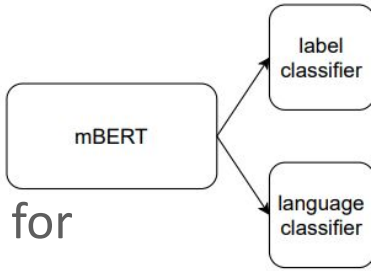
Evidence from domain adaptation:

Gradient reversal helps to encourage the features learned by the model to be **domain-invariant** by training the model to **confuse a domain classifier**

We did the same for language

We encourage the features learned by the model to be **language-invariant** by training the model to confuse a **language identifier**

Language-unlearning with mBERT II: entropy maximisation



Phase 1: the language classification layer is trained in isolation for one epoch on the Wikipedia data set

Phase 2: the target classifier is trained together with mBERT with two goals:
maximise the accuracy of the label classifier and
maximise the entropy of the language classifier's output probabilities

This makes the language classifier approach a **uniform distribution,** which makes the encodings **as confusing as possible** to the language classifier

Results from language-unlearning experiments

	UDPOS		XNLI	
	Init.	Fine-T.	Init.	Fine-T.
Target task	51.2	59.6	29.7	66.3
Lang. ID (Target)	78.3	0.3	49.8	39.2
Lang. ID (Wiki)	59.3	0.5	97.0	97.2

Table 1: Macro F1 scores (%) for target tasks (UDPOS and XNLI) and language identification before (Init.) and after fine-tuning (Fine-T.). Note that 'Lang. ID (Target)' refers to language classification on the target data set.

	UDPOS		XNLI	
	Grad.	Ent.	Grad.	Ent.
Target task	53.5	56.8	62.2	62.1
Lang. ID (Target)	0.1	5.5	1.3	3.4
Lang. ID (Wiki)	0.1	3.1	1.5	54.3

Table 2: Macro F1 scores (%) for target tasks (UDPOS and XNLI) and language identification after training using gradient reversal (Grad.) and entropy maximisation (Ent.). Note that 'Lang. ID (Target)' refers to language classification on the target data set.

Both gradient reversal and entropy maximisation have a negative impact on target task performance

More details in our paper

On the Language-specificity of Multilingual BERT and the Impact of Fine-tuning

Marc Tanti[1] **Lonneke van der Plas**[2] **Claudia Borg**[3] **Albert Gatt**[4]

[1]University of Malta, Institute of Linguistics and Language Technology
[2]Idiap Research Institute
[3]University of Malta, Department of AI
[4]Utrecht University, Information and Computing Sciences
{marc.tanti,claudia.borg}@um.edu.mt
lonneke.vanderplas@idiap.ch, a.gatt@uu.nl

@blackboxNLP 2021

Cross-lingual transfer and computational creativity

Impact of fine-tuning on language-specificity of mBERT

Generating novel concepts

Threats of current AI systems

- Brittleness
- Data-hungriness
- Bias
- Lack of interpretability
- Narrowness

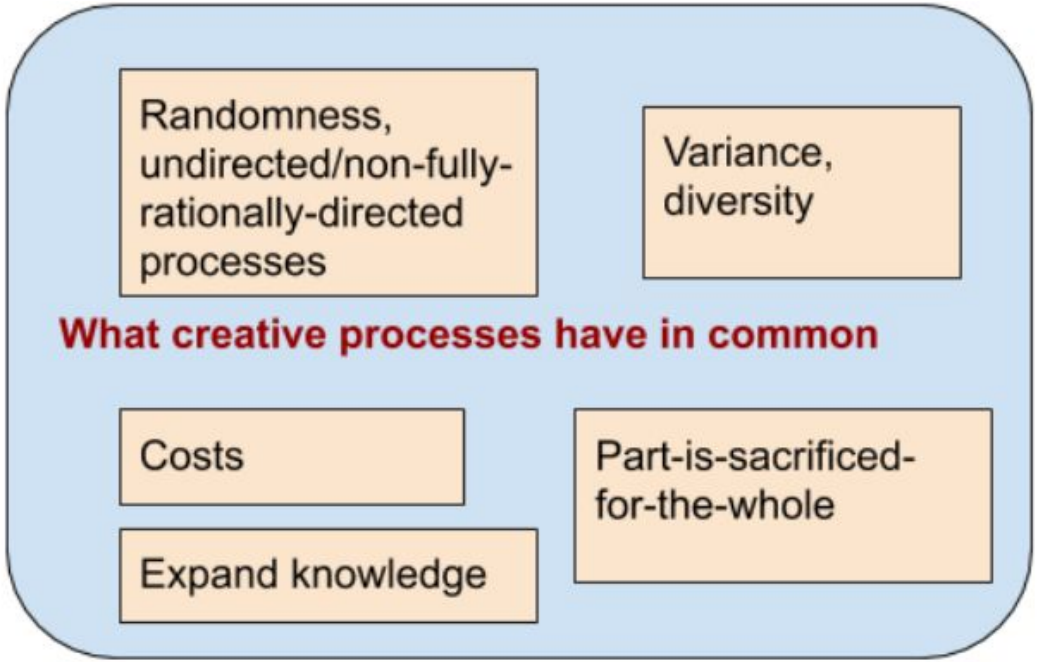
> The threats of the latter have been under-explored

How narrowly defined AI systems threaten society

Work with Michele Loi, during research fellowship DSI Zurich

- Society is governed by processes that allow for diversity and innovation (e.g., market dynamics, natural evolution)
- A society which is highly informed by intelligent systems that are trained in a supervised fashion with a narrowly defined objective functions will not exhibit the same exploration power as a system based on the individuals' judgments
- Fewer agents will be taking over the decision making that was previously done by many more individuals
- More and more impoverished data in training cycle

(Loi & Van der Plas, SDS 2020) (Loi et al., ICCC 2020)



Work that introduces some form of creativity:

- Novelty search (Lehman & Stanley, 2011)
- Intrinsically Motivated Reinforcement Learning (Kaplan & Oudeyer, 2006)
- Work on using off-policy learning for recommender systems to avoid ’myopic recommendations’, where the short term reward overshadows long-term user utility (Ma et al., 2020)

Creative thinking

Creativity, a much needed skill

“We need new ideas to solve our country’s pressing problems”

“We need workers who can *think outside the box* - especially in science and technology - to be competitive in today’s global economy”

[Moran, 2010]

Creativity is most wanted skill according to LinkedIn Learning

[learning.linkedin.com]

AI for creative thinking

Forecast: The global computational creativity market size to grow from USD 204 million in 2018 to USD 685 million by 2023, at a CAGR of 27.4% during 2018–2023

Still, an underexplored topic

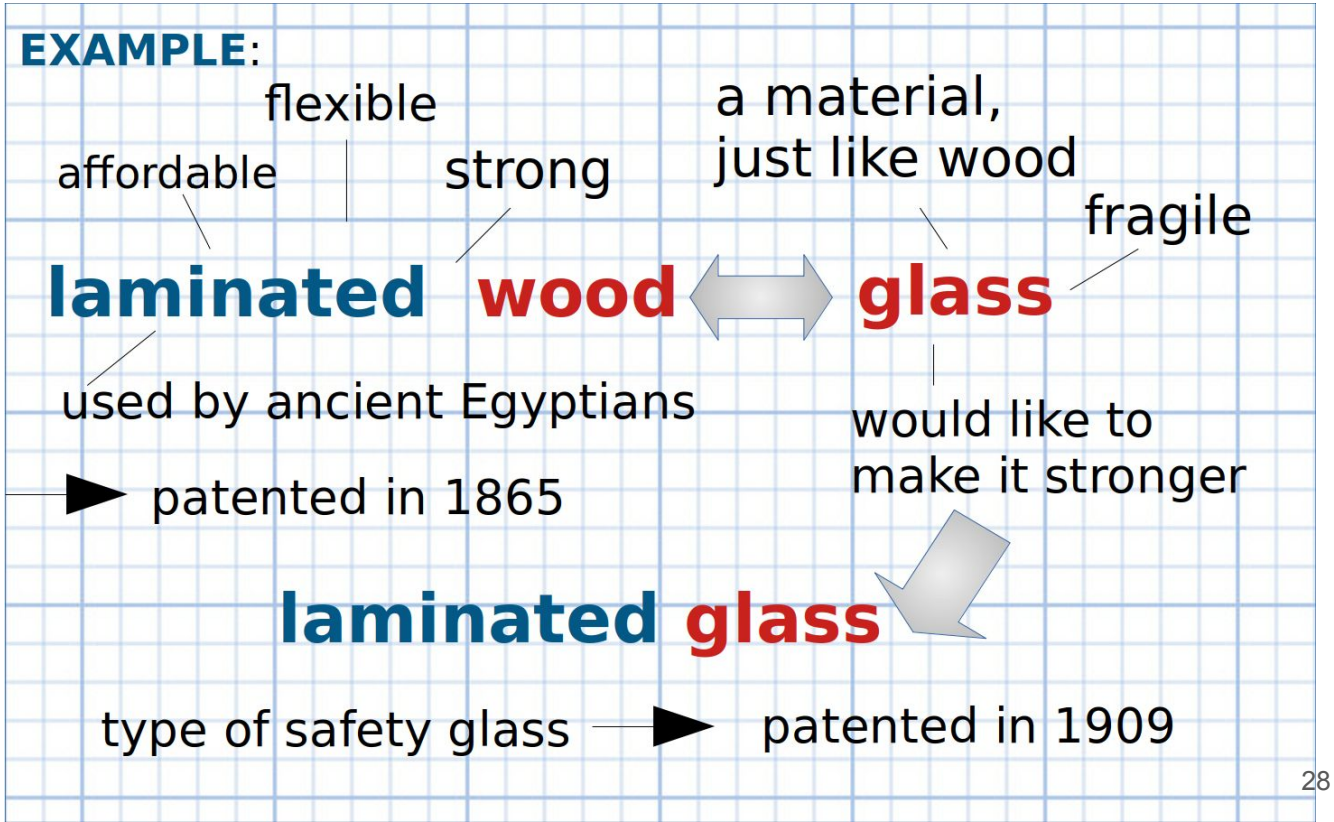
[Source: www.researchandmarkets.com/
Machine Translation to grow only 15%, chatbots 28%]

AI for creative thinking

Creative thinking follows certain patterns

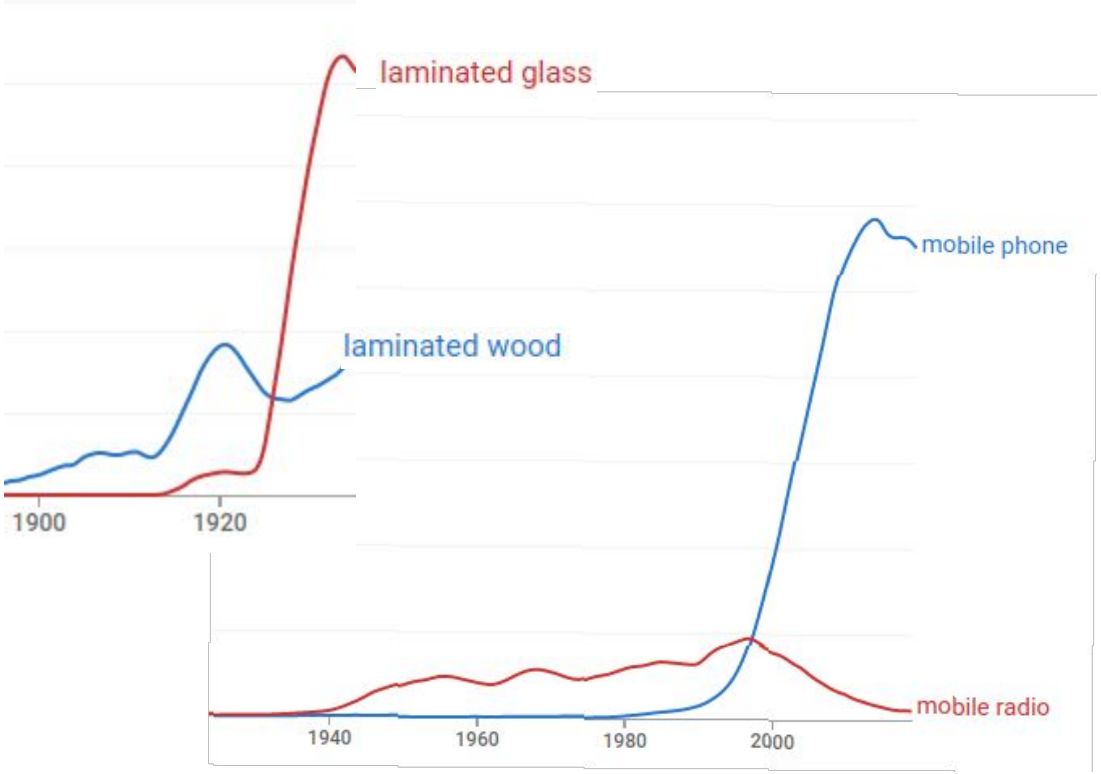
Can be learned by machine

Need to process large amounts of text



AI for creative thinking

We can trace the emergence and success of new ideas in texts



Compounds

Examples: vaccination certificate, flight schedule, stress management, PCR test, quarantine hotel…

- The formation of a new lexeme by adjoining two or more lexemes (Bauer, 2003:40)
- Studied extensively in linguistic literature and more and more attention in the field of Natural Language Processing (NLP)
- Compounding is a very productive word formation process
  - English-speaking children can create novel compounds in spontaneous speech from a very young age (Clark, 1981)
- Very common word type, but many occur with a very low token count

- High productivity makes compositional approaches to automatic processing indispensable

- Also, it raises questions about the processes that underlie the generation of novel compounds

Noun-noun compound interpretation

leather jacket → **jacket *made of* leather ‘veste en cuir’**

leather scissors → **scissors *used to cut* leather ‘ciseaux pour le cuir’**

kitchen knife → **knife *used in the* kitchen ‘couteau de cuisine’**

cheese knife → **knife *used to cut* cheese ‘couteau à fromage’**

Comparing Linguistic and Visuo-Linguistic Representations for Noun-Noun Compound Relation Classification in English

Inga Lang

Supervised by Albert Gatt

Co-supervised by Lonneke van der Plas and Malvina Nissim

Compounds as vehicles for creative thought

(Work in collaboration with Prajit Dhar and Inga lang)

- Compounds allow us to do conceptual recombination
- Using known concepts in combination to create novel ones
- Very flexible, no need to specify the relation between the constituents

Novel compound generation: How?

- Using distributed representations for the constituents and modelling their combination
- For example, glass-bottom boat is found in early corpora, but not glass canoe
- Task: Infer that a glass canoe is a plausible concept, given the evidence of seeing glass-bottom boat, and the similarity between the constituents

Dense vector-based representations

- Condensed space smooths the distribution and adds generalisation power, a remedy for data sparseness
- However, not only does it cater for unseen events (that are likely given the overall distribution)
- It is also an opportunity for the creation of novel (truly unseen) but plausible combinations
- Support from CogSci: semantic networks of low and high semantic creative individuals have different structural properties [Kennet et al, 2014]
- High semantic creative individuals use the same simple search processes to reach further and to more weakly connected concepts [Kennet and Austerweil, 2016]

time-stamped
text data

TRAIN

DEV

EVALUATE

1990 2000 2005 2010 2015 2019

Time

tea
coffee
water

Vector-based
semantic model
of words and
their relations

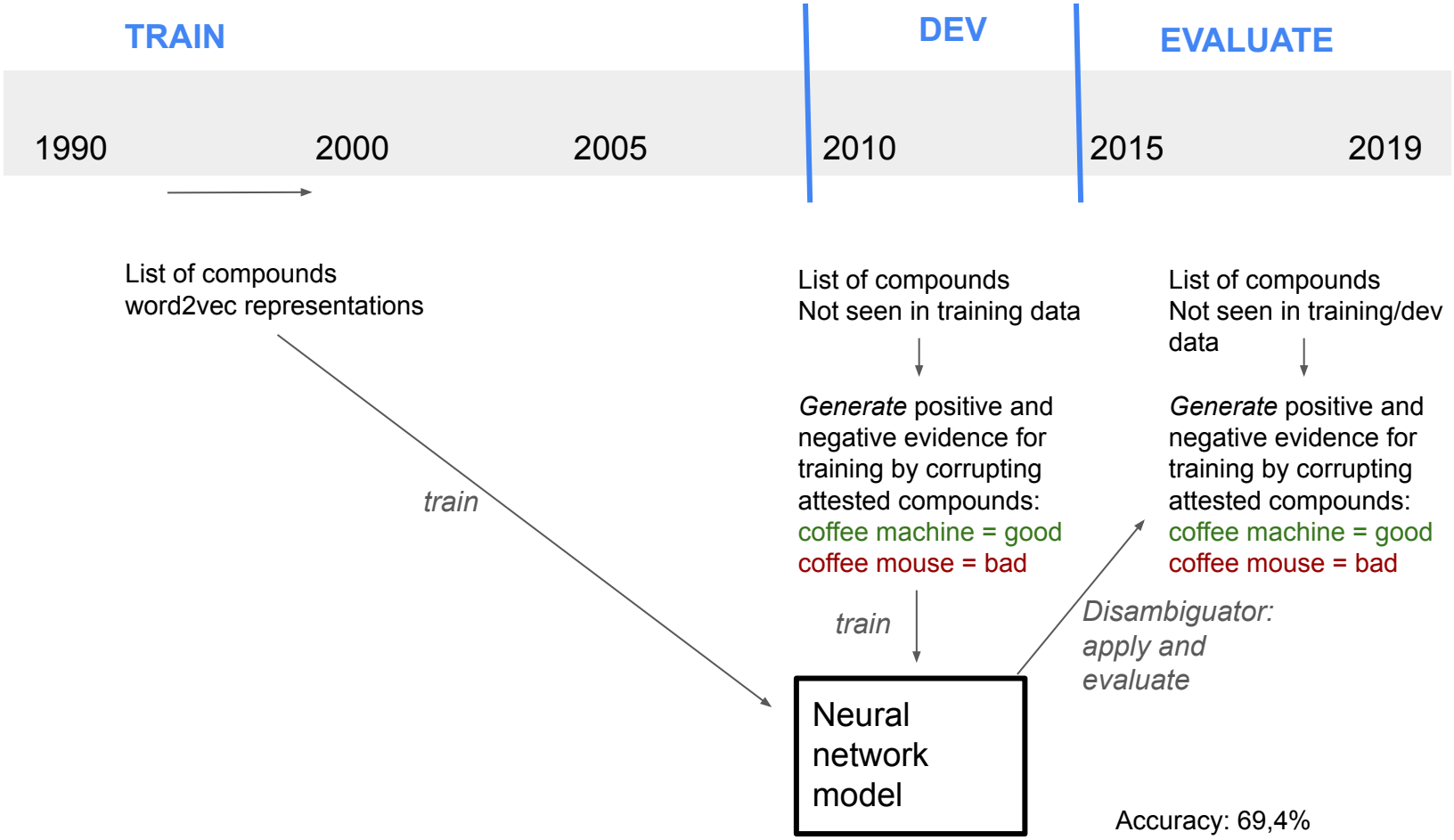
Trends

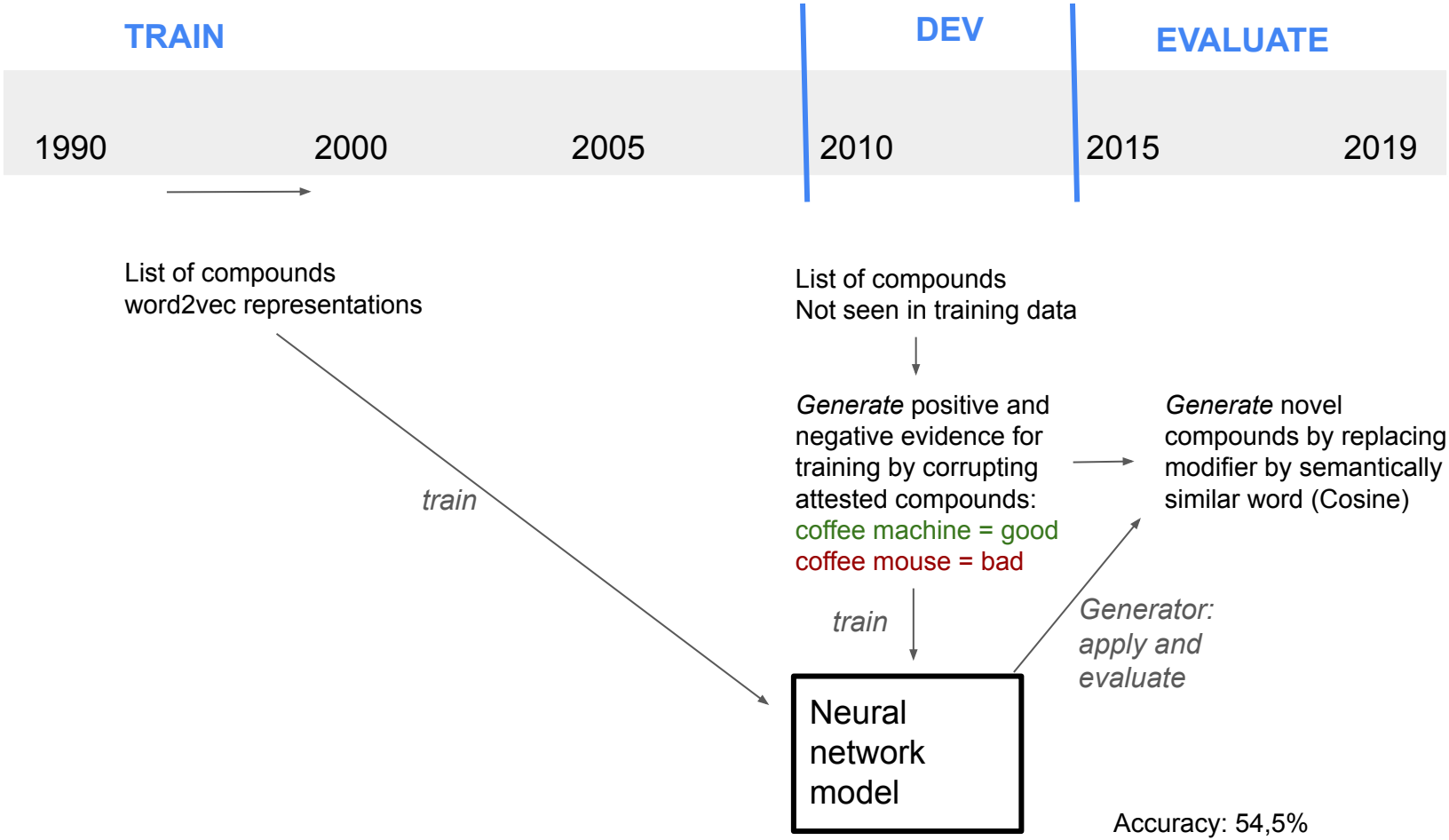
1980 1990

coffee machine = good
coffee mouse = bad

Train model
to learn what are
good combinations

Predict
new ideas

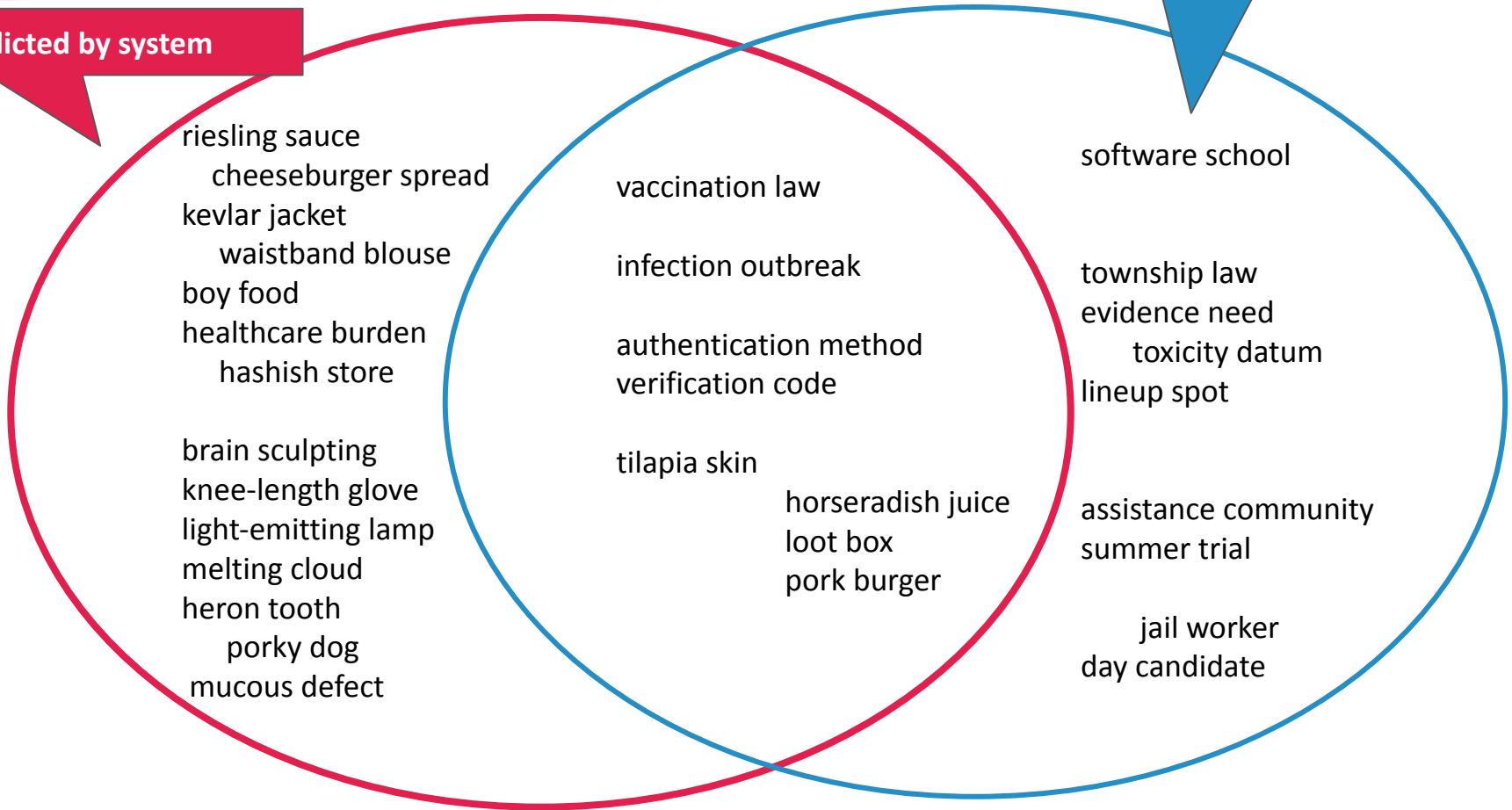

TRAIN
DEV
EVALUATE
1990
2000
2005
2010
2015
2019
List of compounds
word2vec representations
train
List of compounds
Not seen in training data
Generate positive and negative evidence for training by corrupting attested compounds:
coffee machine = good
coffee mouse = bad
train
List of compounds
Not seen in training/dev data
Generate positive and negative evidence for training by corrupting attested compounds:
coffee machine = good
coffee mouse = bad
Disambiguator: apply and evaluate
Neural network model
Accuracy: 69,4%


TRAIN
DEV
EVALUATE
1990
2000
2005
2010
2015
2019
List of compounds
word2vec representations
train
List of compounds
Not seen in training data
Generate positive and negative evidence for training by corrupting attested compounds:
coffee machine = good
coffee mouse = bad
train
Generate novel compounds by replacing modifier by semantically similar word (Cosine)
Generator: apply and evaluate
Neural network model
Accuracy: 54,5%

System output Generator

Predicted by system

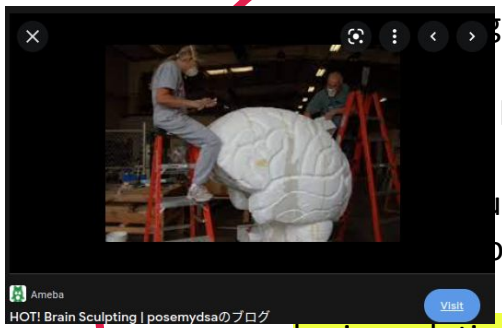
Found in evaluation set 2015-2019



System output Generator

Predicted by system

Found in evaluation set 2015-2019



riesling sauce
ger spread
blouse
urden
ore
brain sculpting
knee-length glove
light-emitting lamp
melting cloud
heron tooth
porky dog
mucous defect

vaccination law
infection outbreak
authentication method
verification code
tilapia skin
horseradish juice
loot box
pork burger

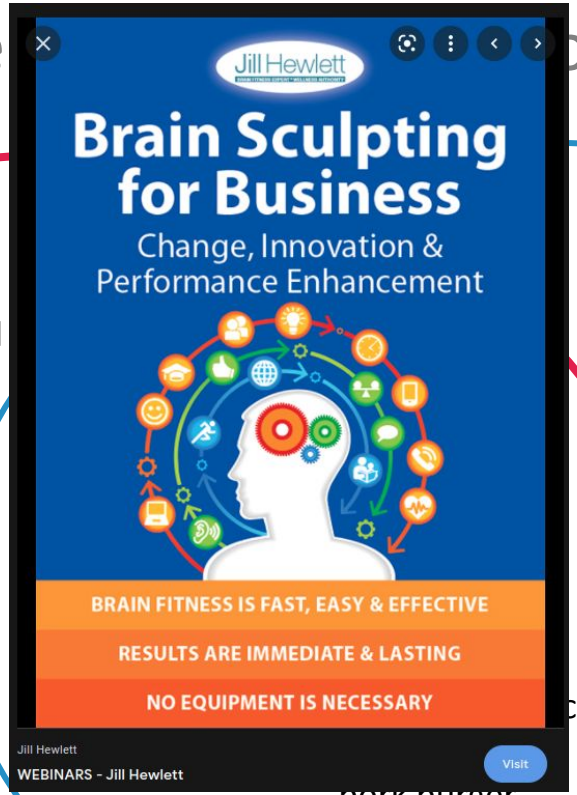
software school
township law
evidence need
toxicity datum
lineup spot
assistance community
summer trial
jail worker
day candidate

Syste or

Predicted by system

riesling sauce
cheeseburger spread
kevlar jacket
waistband blouse
boy food
healthcare burden
hashish store

brain sculpting
knee-length glove
light-emitting lamp
melting cloud
heron tooth
porky dog
mucous defect



Found in evaluation set 2015-2019

software school

township law
evidence need
toxicity datum
lineup spot

assistance community
summer trial

jail worker
day candidate

System output Generator

Predicted by system

riesling sauce
cheeseburger spread
kevlar jacket
waistband blouse
boy food
healthcare burden
hashish store

brain sculpting
knee-length glove
light-emitting lamp
melting cloud
heron tooth
porky dog
mucous defect

Found in evaluation set 2015-2019

software school

ship law
nce need
xicity datum
spot

ance community
er trial

il worker
andidate

System output Generator

Found in evaluation set 2015-2019

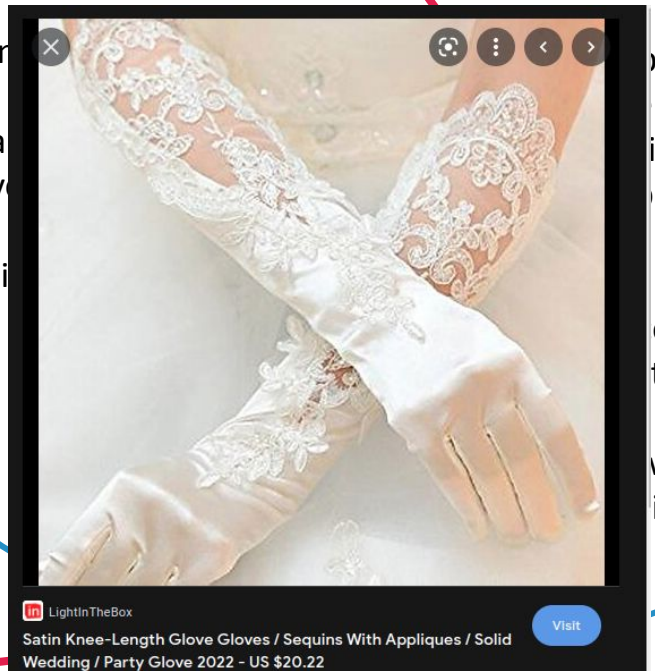
Predicted by system

riesling sauce
cheeseburger spread
kevlar jacket
waistband blouse
boy food
healthcare burden
hashish store

brain sculpting
knee-length glove
light-emitting lamp
melting cloud
heron tooth
porky dog
mucous defect

vaccination law

software school



Evaluation is a challenge

- A number of compounds among our false positives seem to be true positives
- How to determine that automatically?
- We thought of using Google searches (counts)
- However, choosing an appropriate threshold for what we can consider to be a ‘good’ compound is difficult

Evaluation is a challenge

Any ideas on how to evaluate novel compounds are very welcome

- Take a random sample of 100 compounds from our list of false positives
- Then get the number of hits returned by Google when searching for each compound

Threshold: minimum 5000 Google hits
Percentage of ‘correct’ compounds among false positives: 66%
Adjusted accuracy if this were true: 84,48%

Counted as correct:
cash counter
porky dog

Counted as incorrect:
mucous defect
mistreatment
complaint
heron tooth

Threshold: above median
Percentage of ‘correct’ compounds among false positives: 50%
Adjusted accuracy if this were true: 77,20%

Counted as correct:
Snowmobile rental
heel sandal

Counted as incorrect:
midmorning train
porky dog

International Create Challenge '21

time-stamped text data

Scientific articles

Social media data

Company-internal data

TRAIN

EVALUATE

1990 2000 2005 2010 2015 2019

Time

coffee tea water

Vector-based semantic model of words and their relations

Trends

1980 1990

coffee machine = good
coffee mouse = bad

Train model to learn what are good combinations

Predict new ideas

Create an interface that allows for topic- specific browsing

Winning team of ICC'21

Industry partners

Merck Serono Aubonne (pharma)

Beverages and food company

Informants & support

Educational publisher

Information science non-p

ICC mentors

IDIAP technical staff

FoodHack

Contact: lonneke.vanderplas@idiap.ch

Cross-lingual transfer and computational creativity

Impact of fine-tuning on language-specificity of mBERT

- mBERT loses language-specificity during fine-tuning for specific tasks
- Language unlearning does not help

Generating novel concepts

- First results on compound plausibility prediction and generation

Thanks for your attention!