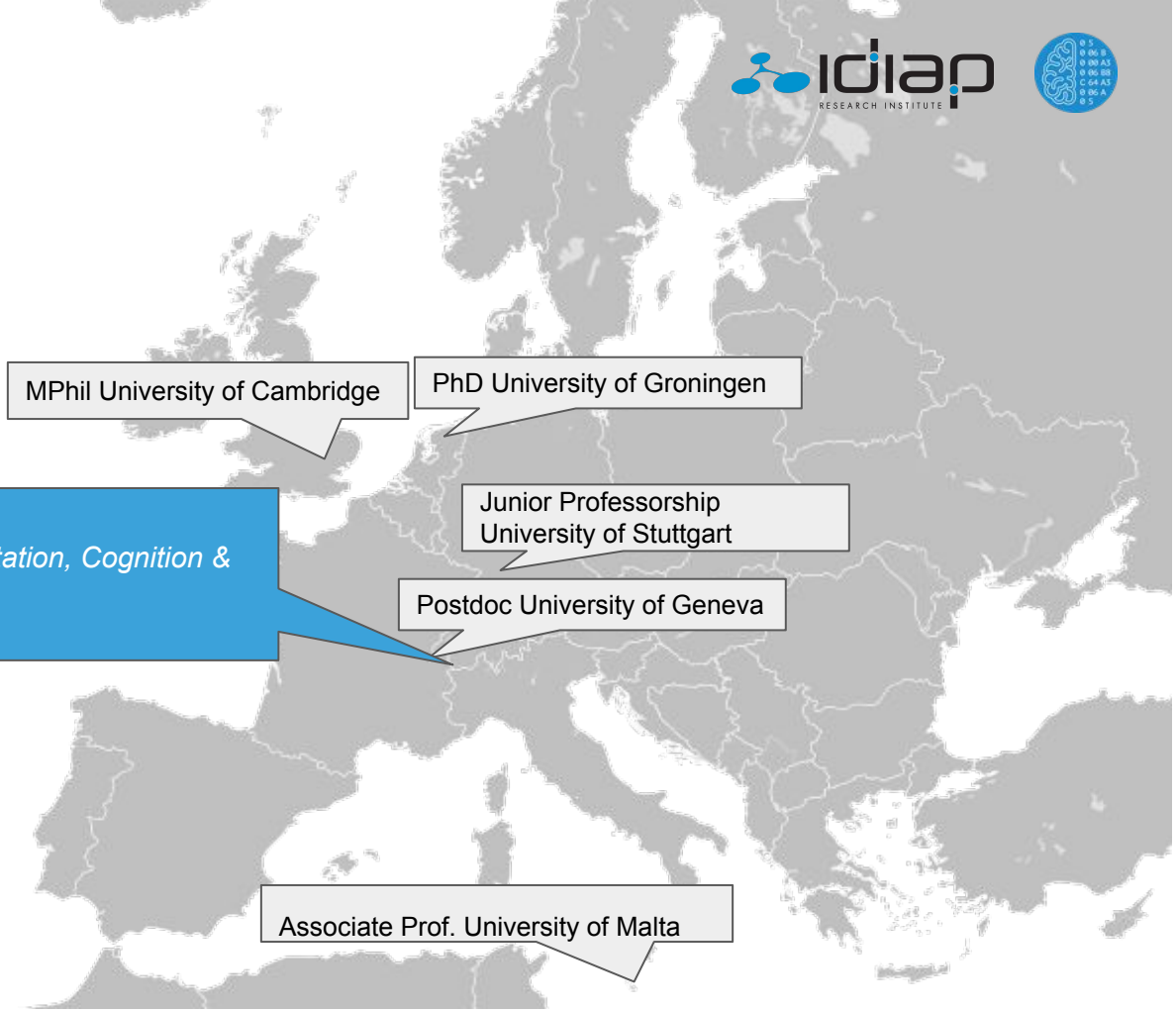
Le traitement du langage naturel en 2023

SSIE workshop, mars 2023

Lonneke van der Plas, Idiap

lonneke.vanderplas@idiap.ch

Qui suis-je?

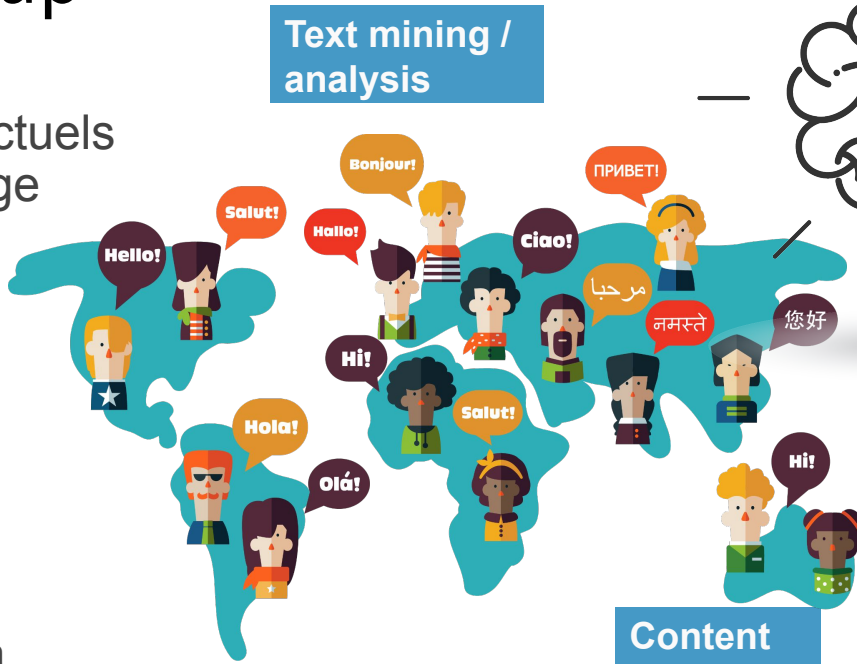


Actuellement responsable du groupe *Computation, Cognition & Language* à l'Idiap à Martigny

Computation, Cognition & Language Group

Limites des systèmes d'IA actuels en ce qui concerne le langage

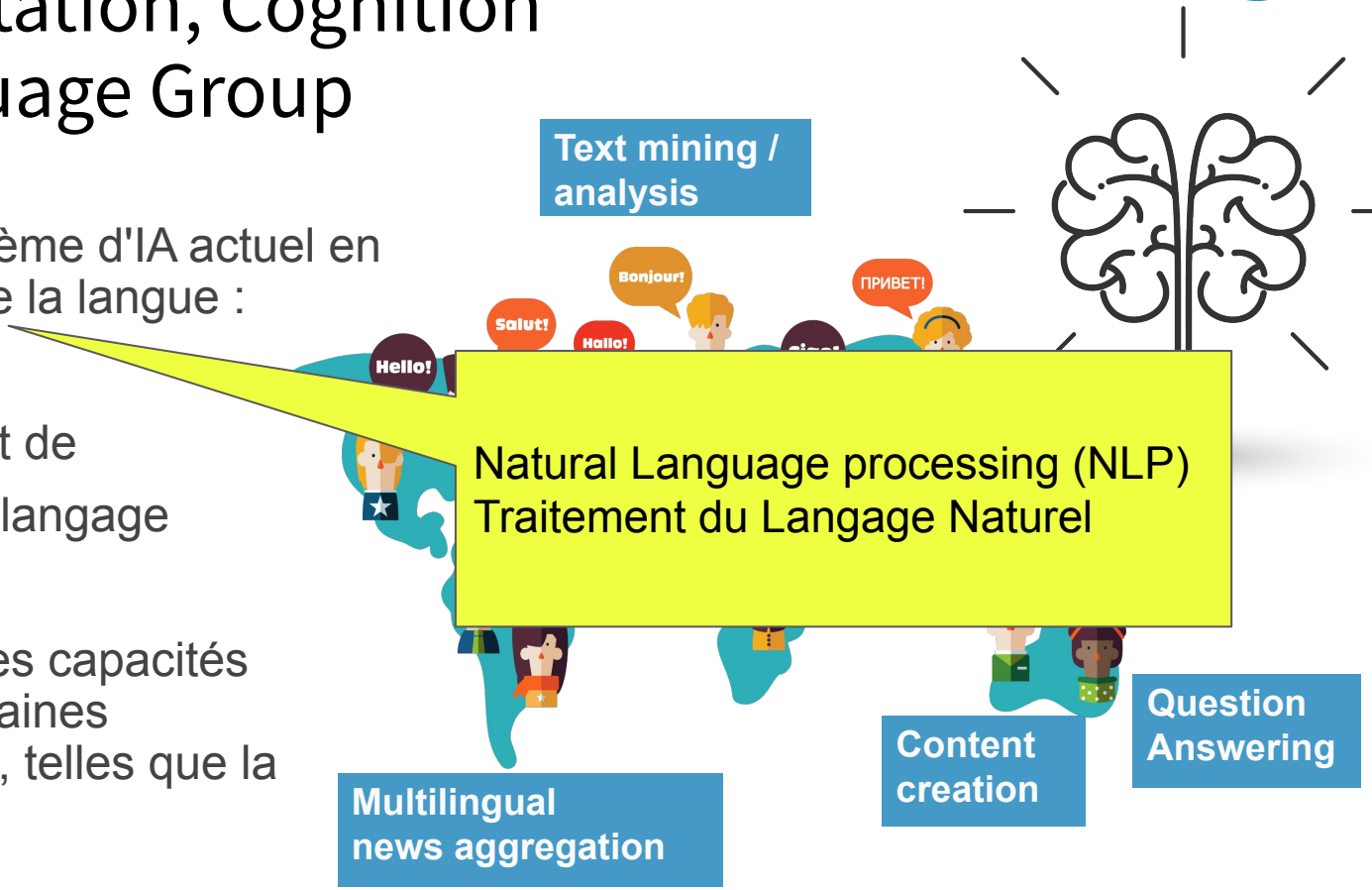
- Développement de technologie du langage multilingue
- Modélisation des capacités cognitives humaines sous-exposées, telles que la créativité



Computation, Cognition & Language Group

- Limites du système d'IA actuel en ce qui concerne la langue :

- Développement de technologie du langage multilingue
- Modélisation des capacités cognitives humaines sous-exposées, telles que la créativité



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

+3 Cross Research Groups

14 research groups

+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

Les merveilles du NLP

- Le domaine du NLP a fait d'énormes progrès
- Il est de plus en plus utilisé dans notre quotidien
- Il y a beaucoup d'avantages, mais aussi plusieurs menaces

Miracle of the Slave by Tintoretto, from Wikimedia Commons

Qu'est-ce que ChatGPT ?

GPT, signifie Generative Pre-trained Transformer

Le "chat" représente la tâche pour laquelle ce modèle GPT a été spécialisé

ChatGPT est un grand modèle informatique de langage naturel

(large language model,LLM)

+ quelques extras

Tous les modèles sont faux, mais certains sont utiles

Que doit contenir un modèle de langage ?

Que doit contenir un modèle de langage ?

Le sens des mots : quels mots sont similaires ?

Que doit contenir un modèle de langage ?

Le sens des mots : quels mots sont similaires ?

Un chien mord un homme ≠ Un homme mord un chien

Que doit contenir un modèle de langage ?

Le sens des mots : quels mots sont similaires ?

La structure du langage

Que doit contenir un modèle de langage ?

Le sens des mots : quels mots sont similaires ?

La structure du langage

Utilisateur : Dois-je apporter un parapluie demain à Martigny ?

Système : Il n'y aura pas de pluie demain à Martigny, donc non.

Que doit contenir un modèle de langage ?

Le sens des mots : quels mots sont similaires ?

La structure du langage

Connaissance du monde

Que doit contenir un modèle de langage ?

Le sens des mots : quels mots sont similaires ?

La structure du langage

Connaissance du monde

Rien d'autre?

Que doit contenir un modèle de langage ?

Le sens des mots : quels mots sont similaires ?

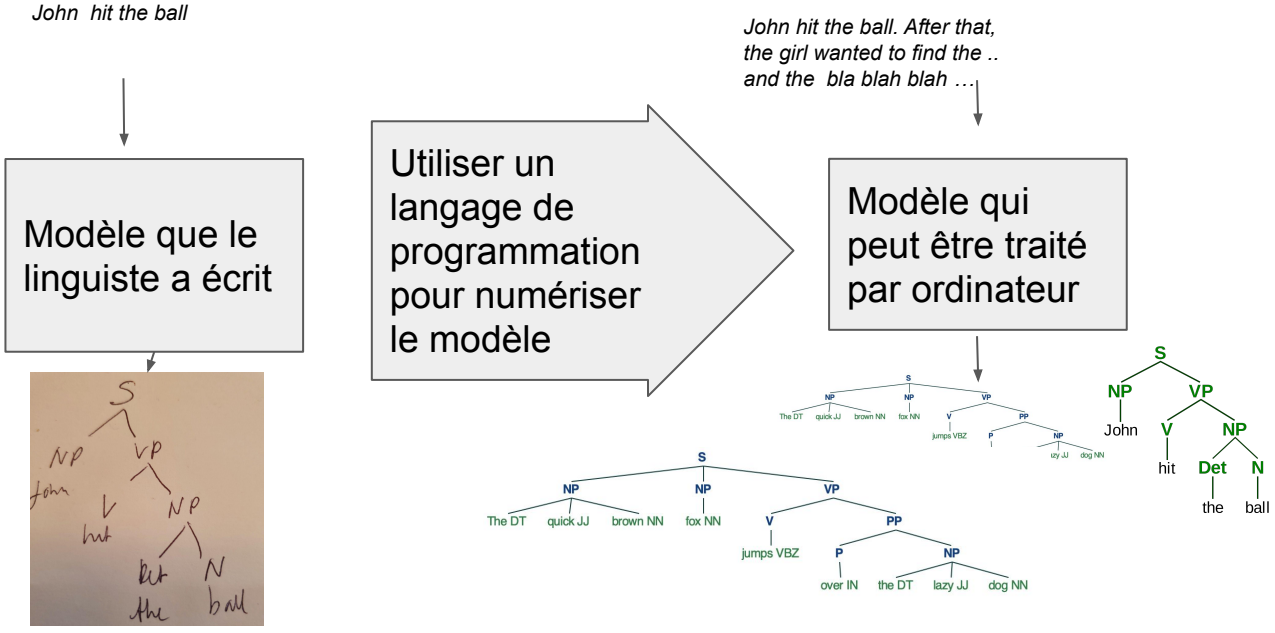
La structure du langage

Connaissance du monde

Savoir raisonner

Modèles informatiques du langage (un peu d'histoire)

Le domaine du NLP/CL a commencé avec des modèles qui étaient des versions informatisées de modèles que les linguistes avaient développés « à l'aide d'un stylo et de papier ».



Modèles informatiques du langage (un peu d'histoire)

Puis vint la révolution statistique… avec un accent sur les données

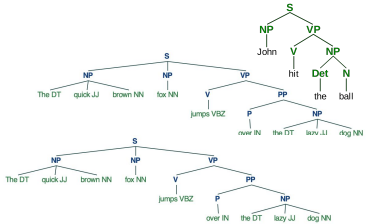
‘Treebanks’ avec des structures/informations linguistiques annotées par des linguistes

Données annotées

Utiliser machine learning pour construire un modèle de langage

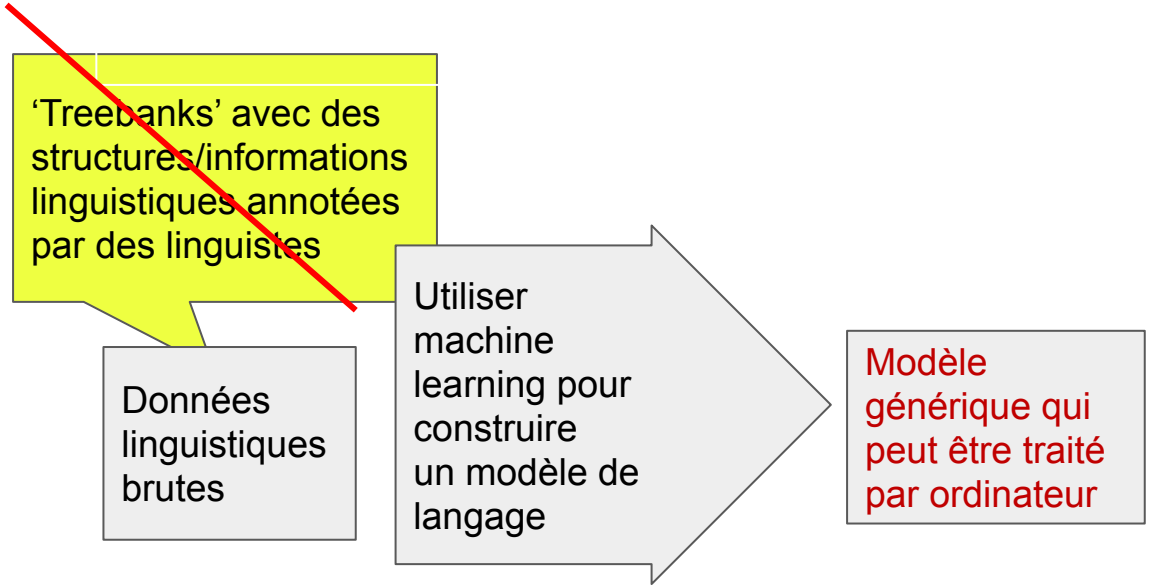
John hit the ball. After that, the girl wanted to find the .. and the bla blah blah …

Modèle générique qui peut être traité par ordinateur



Modèles informatiques du langage (un peu d'histoire)

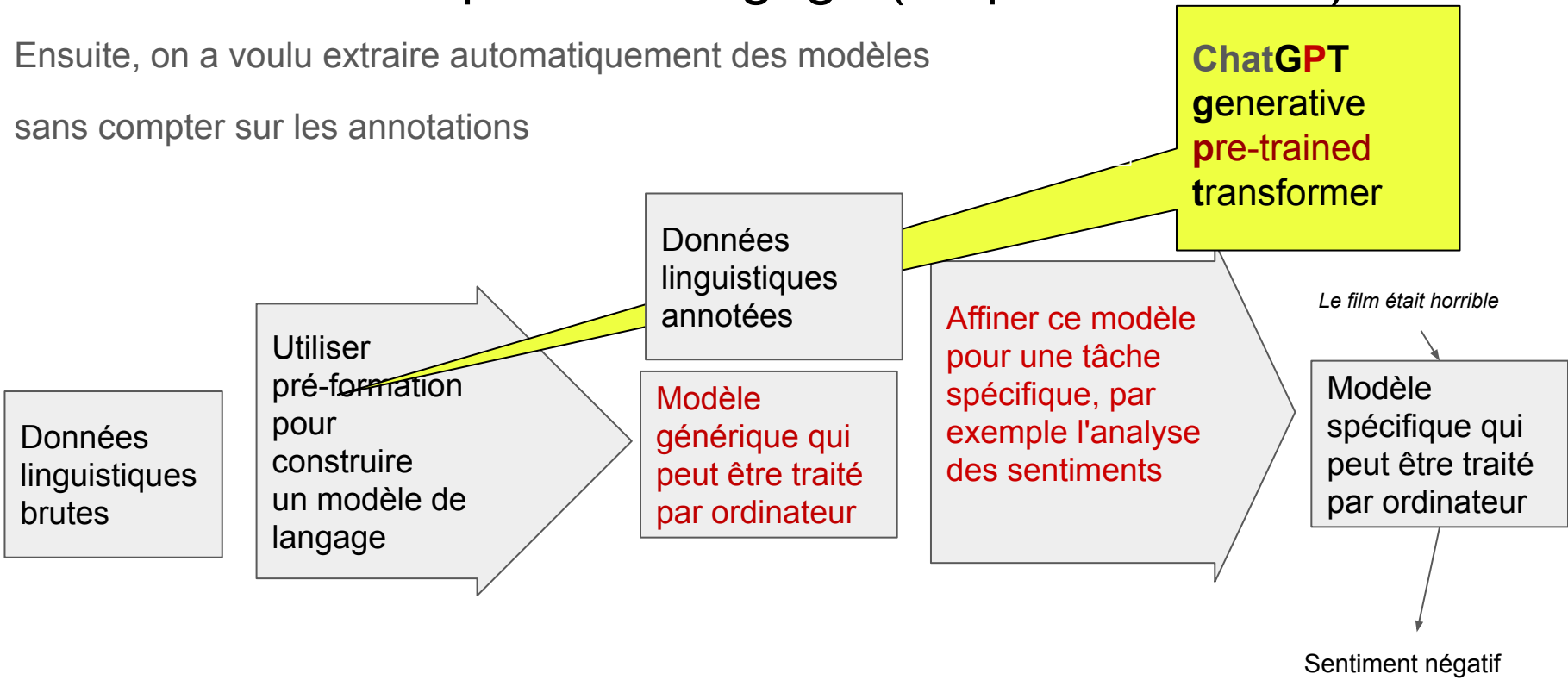
Ensuite, on a voulu extraire automatiquement des modèles sans compter sur les annotations



Modèles informatiques du langage (un peu d'histoire)

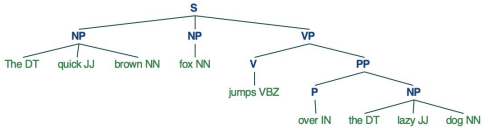
Ensuite, on a voulu extraire automatiquement des modèles

sans compter sur les annotations



Comment les machines apprennent-elles ?

- Ils peuvent apprendre de manière supervisée
  - Ils ont besoin de données annotées/étiquetées en entrée
    - Les annotations sont faites par des humains et coûtent de l'argent
    - Non disponible pour toutes les langues du monde
- Ils peuvent apprendre de manière non supervisée
  - Vous leur donnez des textes bruts et ils devraient comprendre comment les données peuvent être organisées
- Ils peuvent apprendre de manière self-supervisée
  - Pas besoin d'étiquettes. Une partie des données est ce qui doit être prédit. Apprend ce qui est bon ou mauvais à partir des données elles-mêmes



Comment les machines apprennent-elles ?

- Ils peuvent apprendre de manière supervisée
  - Ils ont besoin de données annotées/étiquetées en entrée
    - Les annotations sont faites par des humains et coûtent de l'argent
    - Non disponible pour toutes les langues du monde
- Ils peuvent apprendre de manière non supervisée
  - Vous leur donnez des textes bruts et ils devraient comprendre comment les données peuvent être organisées
- Ils peuvent apprendre de manière self-supervisée
  - Pas besoin d'étiquettes. Une partie des données est ce qui doit être prédit. Apprend ce qui est bon ou mauvais à partir des données elles-mêmes

Ces modèles intègrent les connaissances qui sont explicitement données

Ces modèles apprennent à trouver des modèles dans les données sans avoir vu d'exemples étiquetés

Ils atteignent de très bonnes performances lorsqu'ils sont utilisés pour plusieurs applications

Que doit contenir un modèle de langage ?

Le sens des mots : quels mots sont similaires ?

La structure du langage

Connaissance du monde

Savoir raisonner

Sémantique distributionnelle

L'hypothèse distributionnelle de Harris (1968) :

Les mots qui apparaissent dans les mêmes contextes ont tendance à avoir des significations similaires

Implication pratique :

Nous pouvons trouver des mots similaires en comparant leur distribution dans les contextes

Z. S. Harris. Mathematical structures of language. Wiley, 1968.

Sémantique distributionnelle

Kzvarit ?????

Les contextes :

Deux bouteilles de **Kzvarit**!

J'aime mon **Kzvarit** *on the rocks*.

Il buvait beaucoup de **Kzvarit**, mais il n'était jamais vraiment ivre.

Sémantique distributionnelle

Les mêmes contextes fonctionnent pour:

Martini

Les contextes :

Deux bouteilles de **Martini** !

J'aime mon **Martini** *on the rocks*.

Il buvait beaucoup de **Martini**, mais il n'était jamais vraiment ivre.

Peut-on calculer la similarité sémantique en utilisant des contextes ?

Comptons combien de fois les mots apparaissent dans plusieurs contextes

Et vérifiez ensuite si des mots sémantiquement similaires apparaissent effectivement dans des contextes similaires

Le chat ressemble-t-il plus au chien qu'au couteau ?

Mots cibles

contextes

	get	see	use	hear	eat	kill
knife	51	20	84	0	3	0
cat	52	58	4	4	6	26
dog	115	83	10	42	33	17
boat	59	39	23	4	0	0
cup	98	14	6	2	1	0
pig	12	17	3	2	9	27
banana	11	2	2	0	18	0

Example from Evert, Baroni, Lenci, 2010, Distributional Semantic Models Tutorial at NAACL-HLT 2010

Comment mesurer la similarité ?

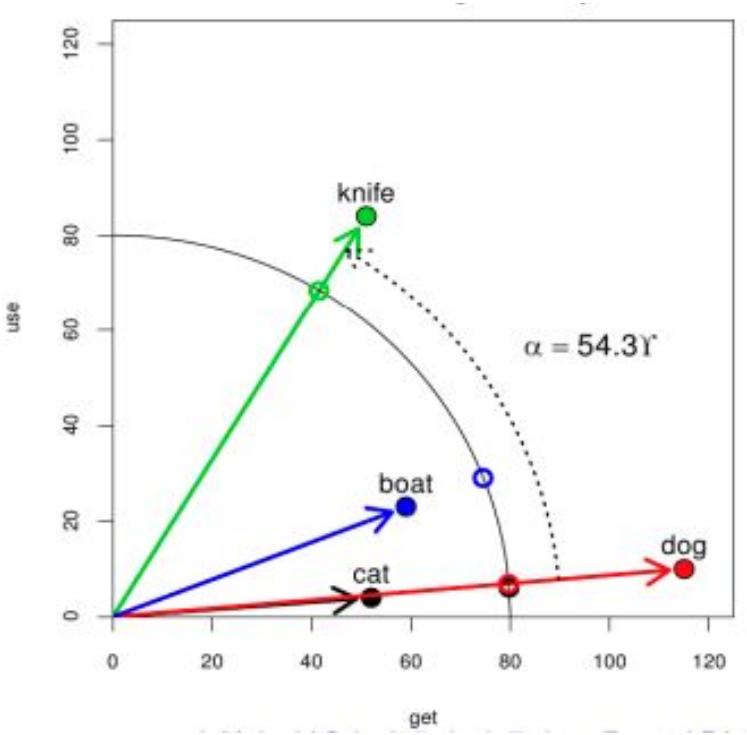
	get	see	use	hear	eat	kill
knife	51	20	84	0	3	0
cat	52	58	4	4	6	26
dog	115	83	10	42	33	17
boat	59	39	23	4	0	0
cup	98	14	6	2	1	0
pig	12	17	3	2	9	27
banana	11	2	2	0	18	0

Mots cibles comme vecteurs

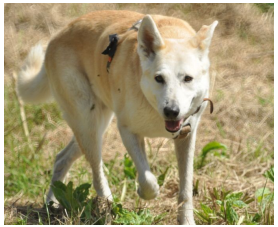
Contextes en tant que dimensions (exemple avec seulement deux contextes/dimensions)

Calculer l'angle entre les vecteurs à l'aide de la mesure cosinus

Vous donne une valeur entre 0 et 1



Caractéristiques sémantiques redéfinies



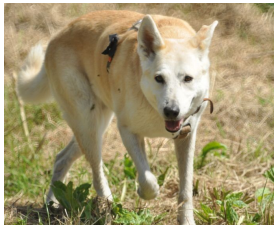
~~[A quatre pattes, est poilu, est vivant]~~

~~[A quatre pattes, n’est pas vivant, est en bois]~~

Photos de Marko Milivojevic - Wikipédia, l'encyclopédie libre

Caractéristiques sémantiques redéfinies

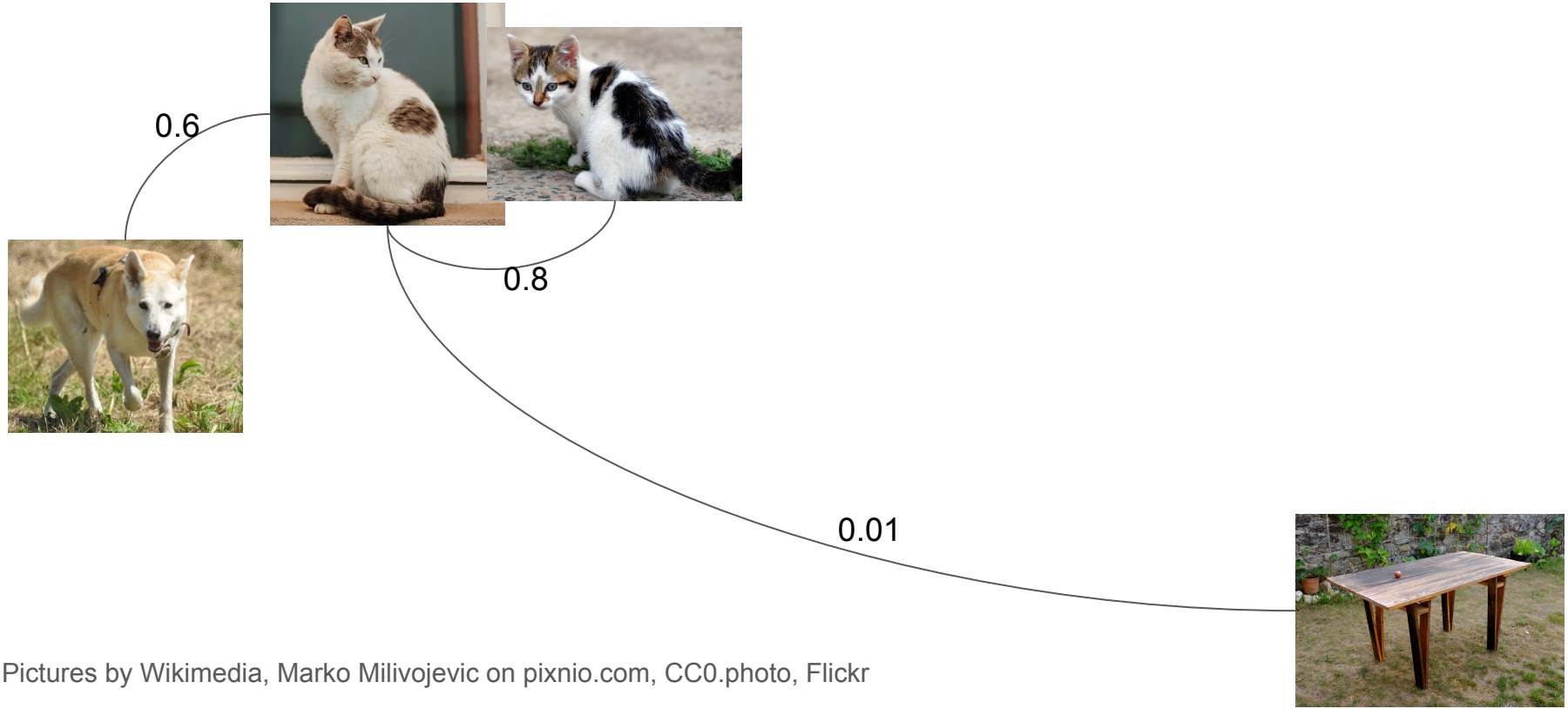
[get: 110,
feed:30,
paint:1
…]



[get: 110,
feed:0,
paint:50
…]

Pictures by Wikimedia, Marko Milivojevic on pixnio.com, CC0.photo, Flickr

Similitude sémantique sur une échelle continue



Pictures by Wikimedia, Marko Milivojevic on pixnio.com, CC0.photo, Flickr

SVD / LSA

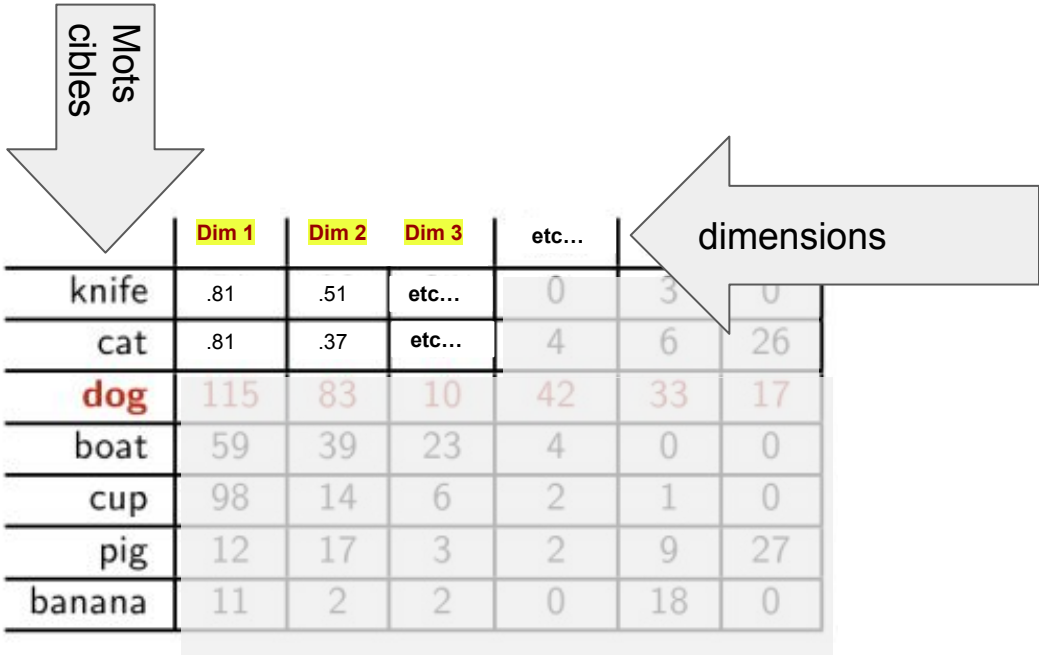
Réduction de la dimensionnalité

En réalité beaucoup plus de contextes, et beaucoup plus de zéros dans les cellules

3	1	0	0	0	0	0	56	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	11	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	51	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	5	0	0	0	0	0	4	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Word2vec, une méthode neuronale auto-supervisée pour obtenir des vecteurs denses

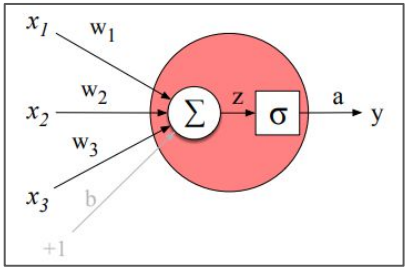
- Software package appelé word2vec (Mikolov et al. 2013a, Mikolov et al. 2013b)
- Le résultat de la formation est un vecteur dense pour chaque mot
- Fonctionne mieux pour chaque tâche NLP que les vecteurs ‘sparse’



	Dim 1	Dim 2	Dim 3	etc…		
knife	.81	.51	etc…	0	3	0
cat	.81	.37	etc…	4	6	26
dog	115	83	10	42	33	17
boat	59	39	23	4	0	0
cup	98	14	6	2	1	0
pig	12	17	3	2	9	27
banana	11	2	2	0	18	0

Exemple de Evert, Baroni, Lenci, 2010, Distributional Semantic Models Tutorial at NAACL-HLT 2010

Modèles de réseaux de neurones artificiels



Les origines résident dans un modèle simplifié du neurone humain et de ses interactions

Aujourd'hui juste un réseau de petites unités de calcul

Ceux-ci prennent un vecteur de valeurs d'entrée et produisent une sortie unique

Ils sont puissants en raison des couches alimentant les couches suivantes

Apprentissage « profond » car les réseaux sont souvent profonds (couches multiples)

Ils apprennent à induire des caractéristiques *(features)* dans le cadre du processus d'apprentissage

Ils nous permettent de faire de la modélisation de bout-en-bout (pas besoin de composants de modèle)

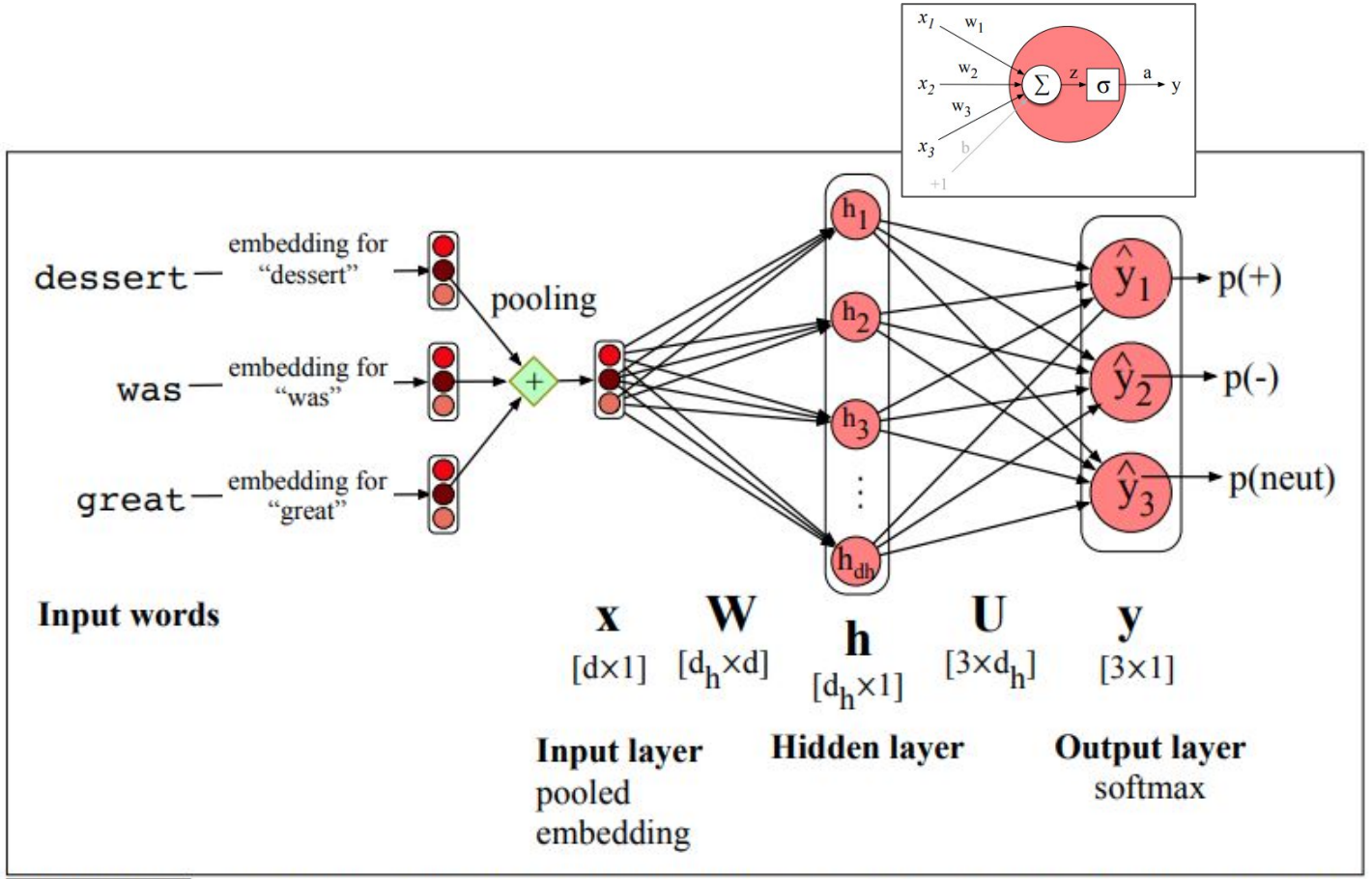


Figure 7.11 Feedforward sentiment analysis using a pooled embedding of the input words.

Picture taken from Speech and Language Processing (3rd ed. draft). 2021. Dan Jurafsky and James H. Martin

Comment fonctionne word2vec ?

Auto-surveillance !

Un des algorithmes : skip-gram avec échantillonnage négatif (SGNS)

Entraîne un classifieur sur une tâche de prédiction :

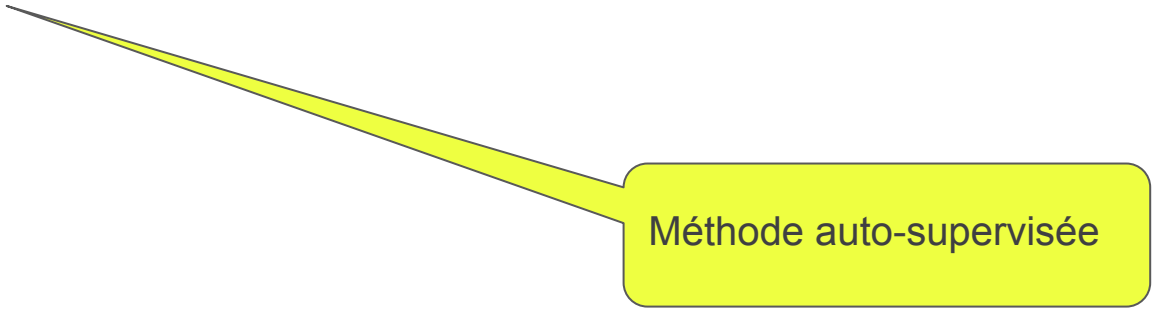
"Le mot w est-il susceptible d'apparaître dans le contexte c ?"

En savoir plus sur word2vec

Entraîne un classifieur sur une tâche de prédiction :

"Le mot w est-il susceptible d'apparaître dans le contexte c ?"

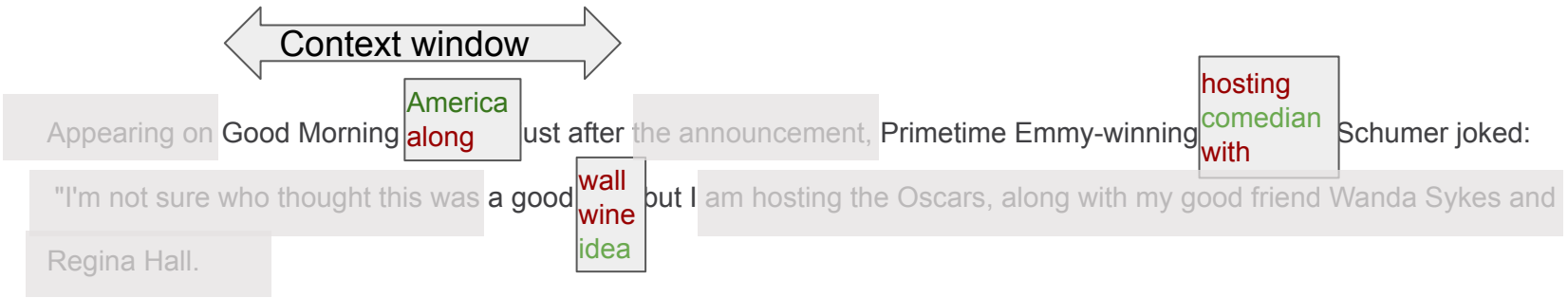
- Exemples positifs : texte réel
- Exemples négatifs : des mots échantillonnés au hasard dans le reste du texte



Exemple

Positive examples

Negative examples



Mots dans le contexte

Paper jam vs strawberry jam

Nous avons parlé des représentations de mots

Mais les mots ont des sens différents selon le contexte

La langue ne se limite pas aux mots. Ce sont des blocs de construction pour de plus longues portions de texte

> Nous avons besoin de modèles qui capturent les relations entre les mots même lorsqu'ils sont éloignés, et nous avons besoin de représentations contextualisées

Les transformateurs sont un type de modèle qui tente de résoudre ce problème

Generative
Pre-trained
Transformer
ChatGPT

Transformeurs (BERT, Devlin et al. 2018)

La modélisation du langage masqué est la tâche d'auto-supervision utilisée pour entraîner le modèle + la prédiction de la phrase suivante

Représentations contextualisées grâce au mécanisme de l'attention

L'attention multi-tête permet au modèle de capturer les multiples relations qui existent entre les mots dans un contexte donné, même lorsqu'ils sont éloignés

Des millions de paramètres entraînables en nombre de couches (base BERT : 12 couches contenant 12 têtes d'attention)

Nécessite d'énormes quantités de données et de calculs (3,3 milliards d'éléments de texte sans étiquette)

Généralement formé une fois (*pre-training*) et utilisé pour plusieurs tâches à l'aide d'une configuration de *fine-tuning*

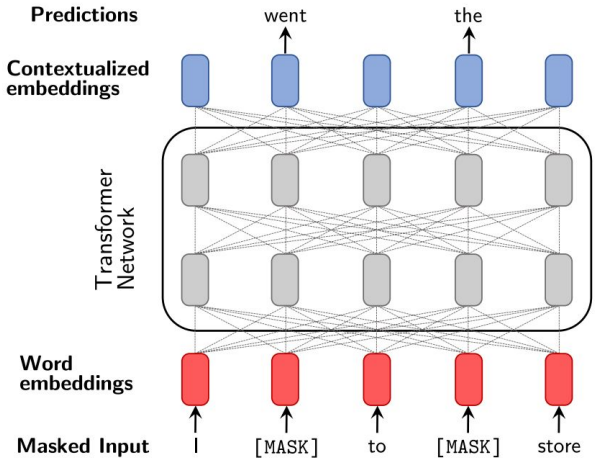
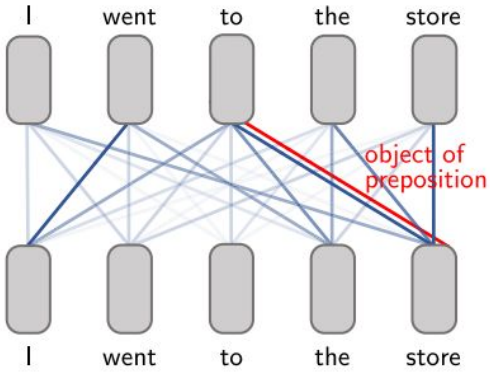


Photo de Christopher D. Manning et al., 2020



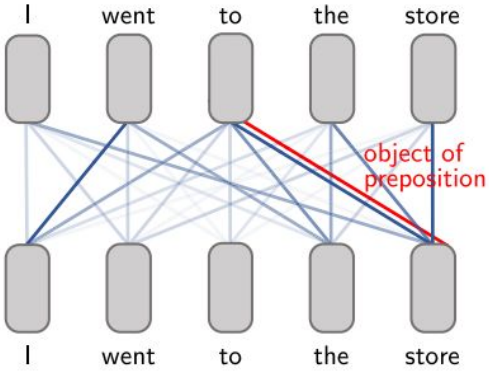
Qu'apprennent les transformeurs ?

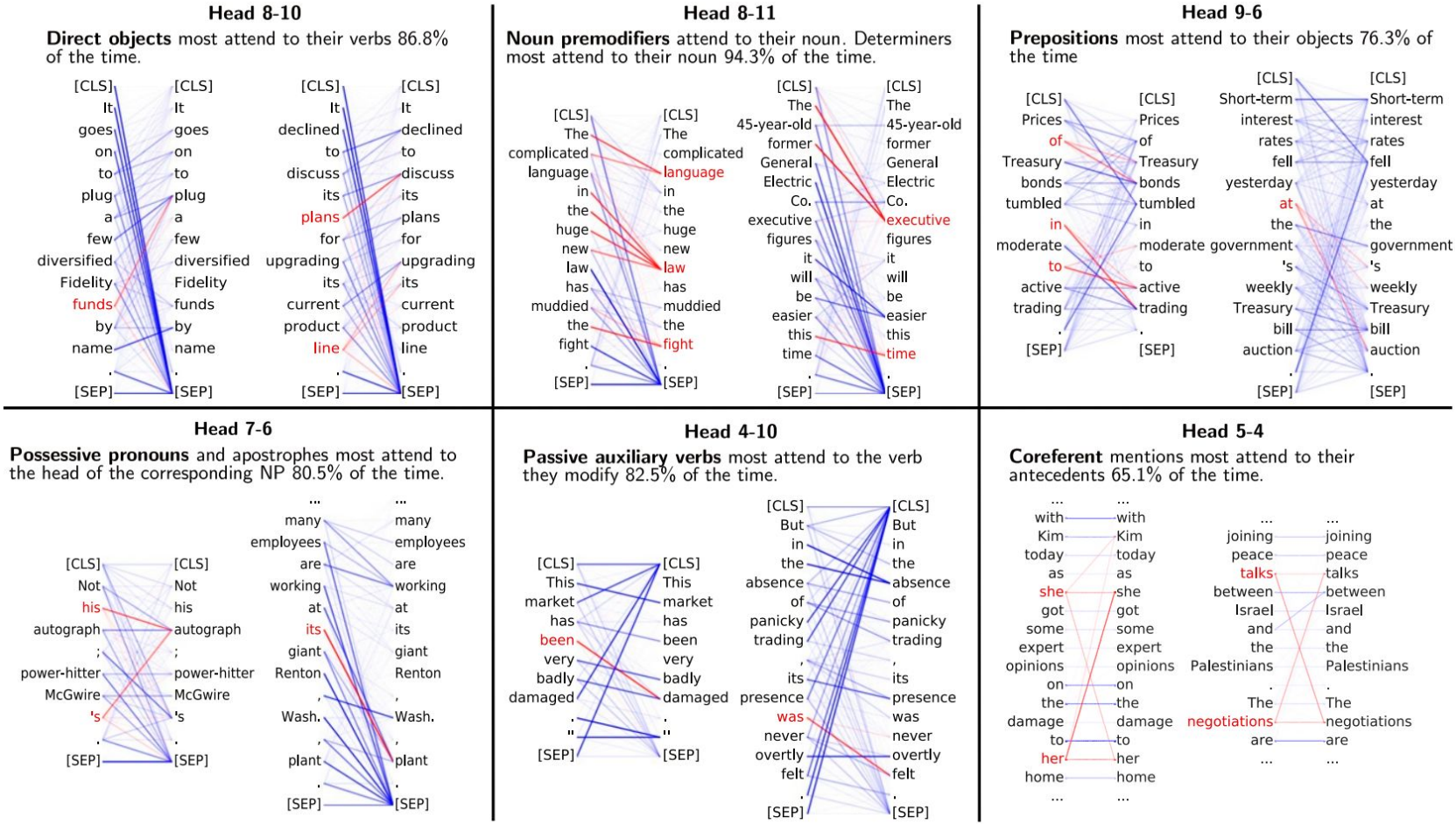
Les têtes d'attention sont destinées à capturer les relations entre les mots

Nous pouvons examiner le mot le plus suivi à chaque position dans une séquence de mots

Cela nous permet de vérifier quel mot de la séquence est le plus "important" pour que le modèle détermine la représentation d'un mot donné

Nous pouvons jeter un œil à l'intérieur des couches spécifiques

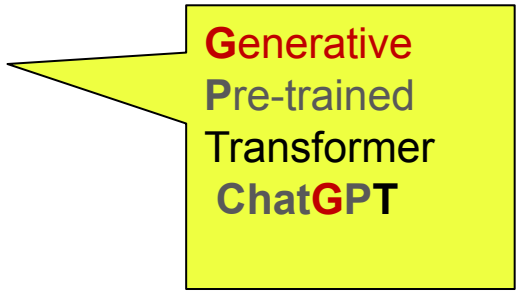




No single head in a single layer corresponds well to dependency syntax overall

Picture from Clark et al., 2019

"Generative"??



Les architectures de transformeurs peuvent être constituées d'un codeur, d'un décodeur ou de l'un de ces deux

Le transformeur BERT est un modèle de codeur

GPT-3 est un décodeur

Il génère du texte en se basant sur un *prompt*

La famille des modèles de langage et leurs défauts

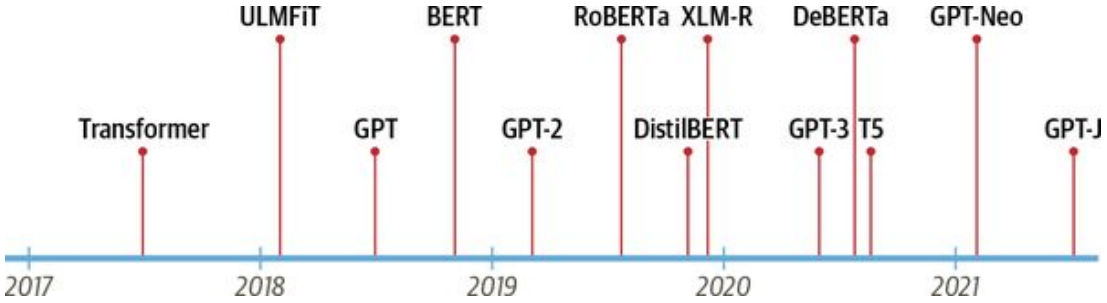
Capturent les patterns statistiques dans le langage

Ont besoin de très grandes quantités de données textuelles et de ressources informatiques massives. Nous n'avons pas accès à ces données

Seuls les «acteurs majeurs» peuvent créer de tels modèles. Ils ne sont pas disponibles pour toutes les langues

Ne savent pas ce qui est vrai ou faux

Ils ne font qu'imiter les textes vus



Panel lors de l'événement chatGPT à l'Idiap le 10 mars

Securité

Al Brown, Fujitsu

Loi

Ana Andrijevic, Centre for Law, Technology and Society

Education

Biljana Petreska von Ritter-Zahony, HEP Vaud

Bertrand Audrin, EHL (HR)

ressources humaines

Le système comme un allié

Attention à la situation de monopole

Attention à ce que vous tapez

Nouvelles compétences en IA nécessaires

Education

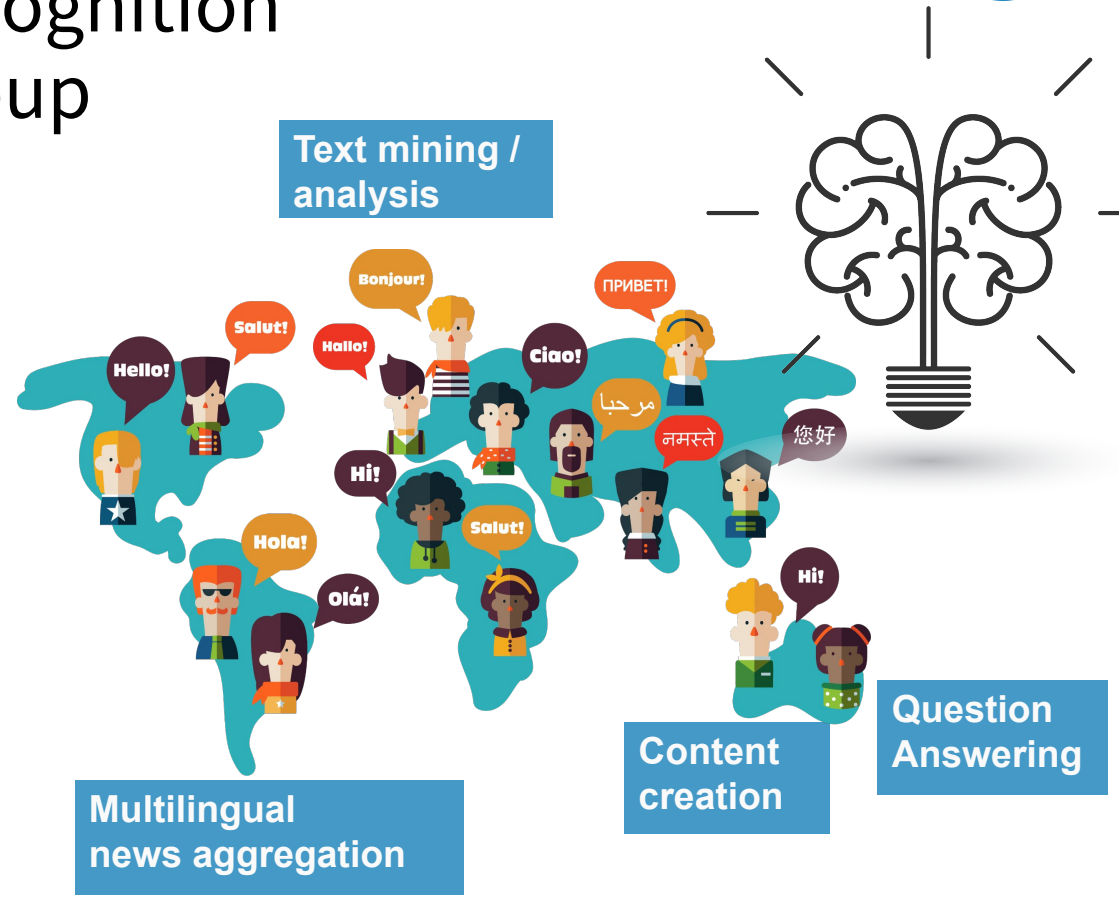
Sensibilisation

éthique

Computation, Cognition & Language Group

Limites du système d'IA actuel en ce qui concerne le langage

- Développement de technologie du langage multilingue
- Modélisation des capacités cognitives humaines sous-exposées, telles que la créativité



l’IA, une menace pour la diversité

La société est régie par des processus qui permettent la diversité et l'innovation

Et si les systèmes intelligents prennent les décisions à notre place ?

Ces processus n’auront pas le même pouvoir d’exploration qu’un système basé sur le jugement des individus

Moins de systèmes prendront les décision, auparavant effectuées par beaucoup plus d'individus

Des données de plus en plus appauvries en cycle de formation

Filter bubbles and echo chambers

Bias in automatic candidate selection

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

Systèmes de recommandation de news

- Les gens lisent souvent les actualités sur leur téléphone portable
- Les flux d'actualités des systèmes de recommandation (tels que YouTube et des médias sociaux) influencent la façon dont les gens découvrent les informations
- Ces flux d'actualités peuvent être personnalisés de manière algorithmique pour chaque utilisateur
- La personnalisation est une fonctionnalité utile
- Cependant, elle limite l'exposition des lecteurs à différents types d'informations

Certaines menaces attribuées aux systèmes de recommandation

- Filter bubble (Pariser 2011)
- Echo chamber (Flaxman et al. 2016)
- Polarization (Dandekar et al. 2013).
- Fragmentation (Helberger 2019)
- Dehumanization (Page et al. 2018)
- ...

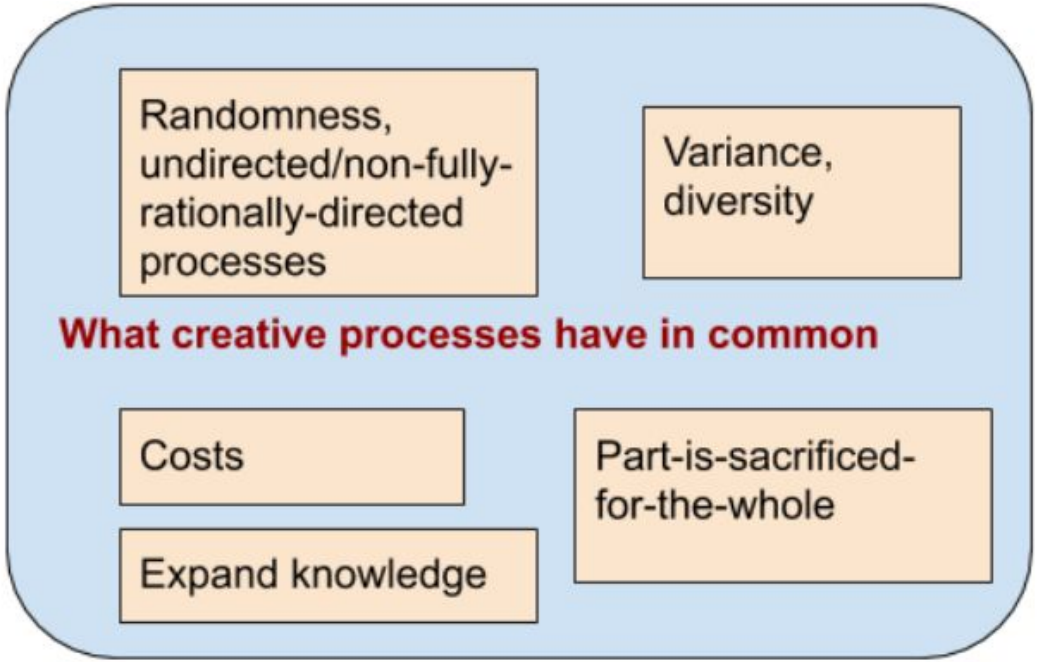
Raza, S., Ding, C. News recommender system: a review of recent progress, challenges, and opportunities. 2022. Artif Intell Rev 55, pages 749–800.

Comment les chercheurs de NRS tentent d'atténuer

Algorithmes sensibles à la diversité

Nudging

Compromis entre diverses mesures d'évaluation



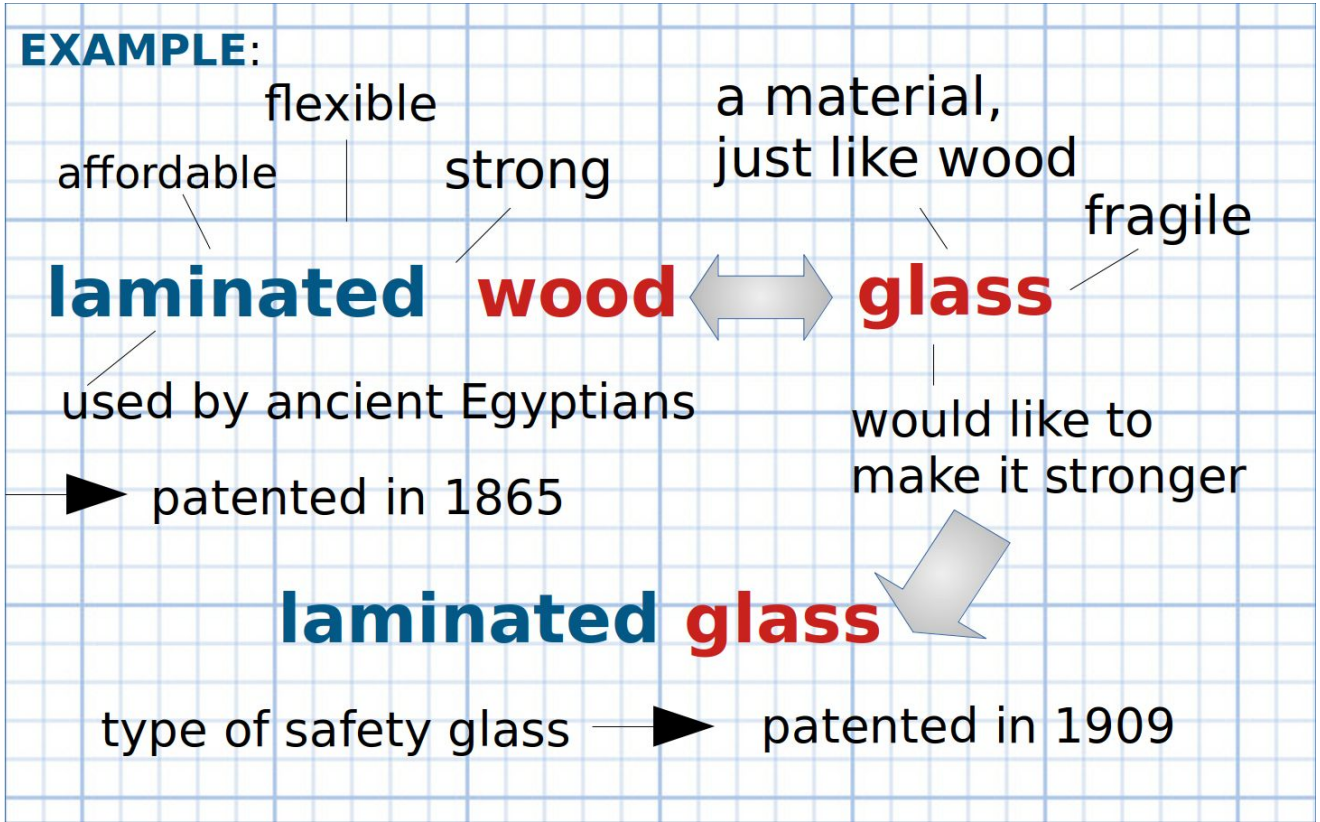
Quelques travaux qui introduisent certains aspects de créativité

- 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)

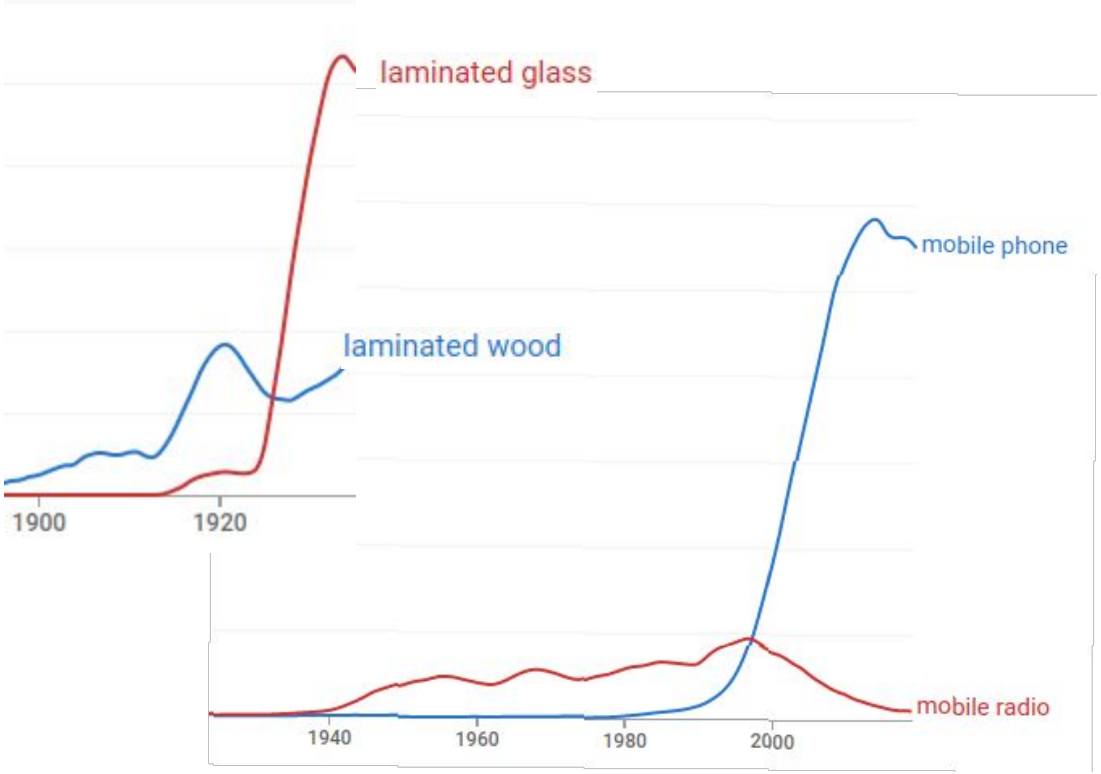
Patterns dans la génération de concepts

La pensée créative suit certains schémas

Peuvent être appris par un système



Emergence et succès de nouveaux concepts dans les corpus



Les mots composés comme véhicules de la pensée créative

- Les mots composés nous permettent de faire de la recombinaison conceptuelle
- Utilisation de concepts connus en combinaison pour en créer de nouveaux
- Les enfants sont capables de produire des mots composés

  (Example: *moon cheese* ...)

- Très flexible, pas besoin de préciser la relation entre les constituants

Génération de nouveaux mots composés : comment ?

- Utiliser des représentations distribuées pour les deux parties et modéliser leur combinaison
- Par exemple, le bateau à fond de verre est connu, mais pas le canoë de verre
- Tâche : Déduire qu'un canoë de verre est un concept plausible

Représentations vectorielles denses

- L'espace condensé lisse la distribution et ajoute la généralisation, remède à la rareté des données
- Cependant, non seulement il s'occupe d'événements rares
- C'est aussi l'occasion de créer des combinaisons inédites (vraiment inédites) mais plausibles
- CogSci : les réseaux sémantiques d'individus de faible créativité et de créativité élevée ont des propriétés structurelles différentes [Kennet et al, 2014]
- Les derniers peuvent atteindre des concepts plus éloignés et plus faiblement connectés plus facilement [Kennet et Austerweil, 2016]

time-stamped
text data

TRAIN

DEV

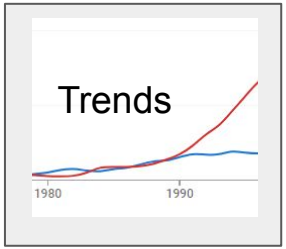
EVALUATE

1990 2000 2005 2010 2015 2019

Time

tea
coffee
water

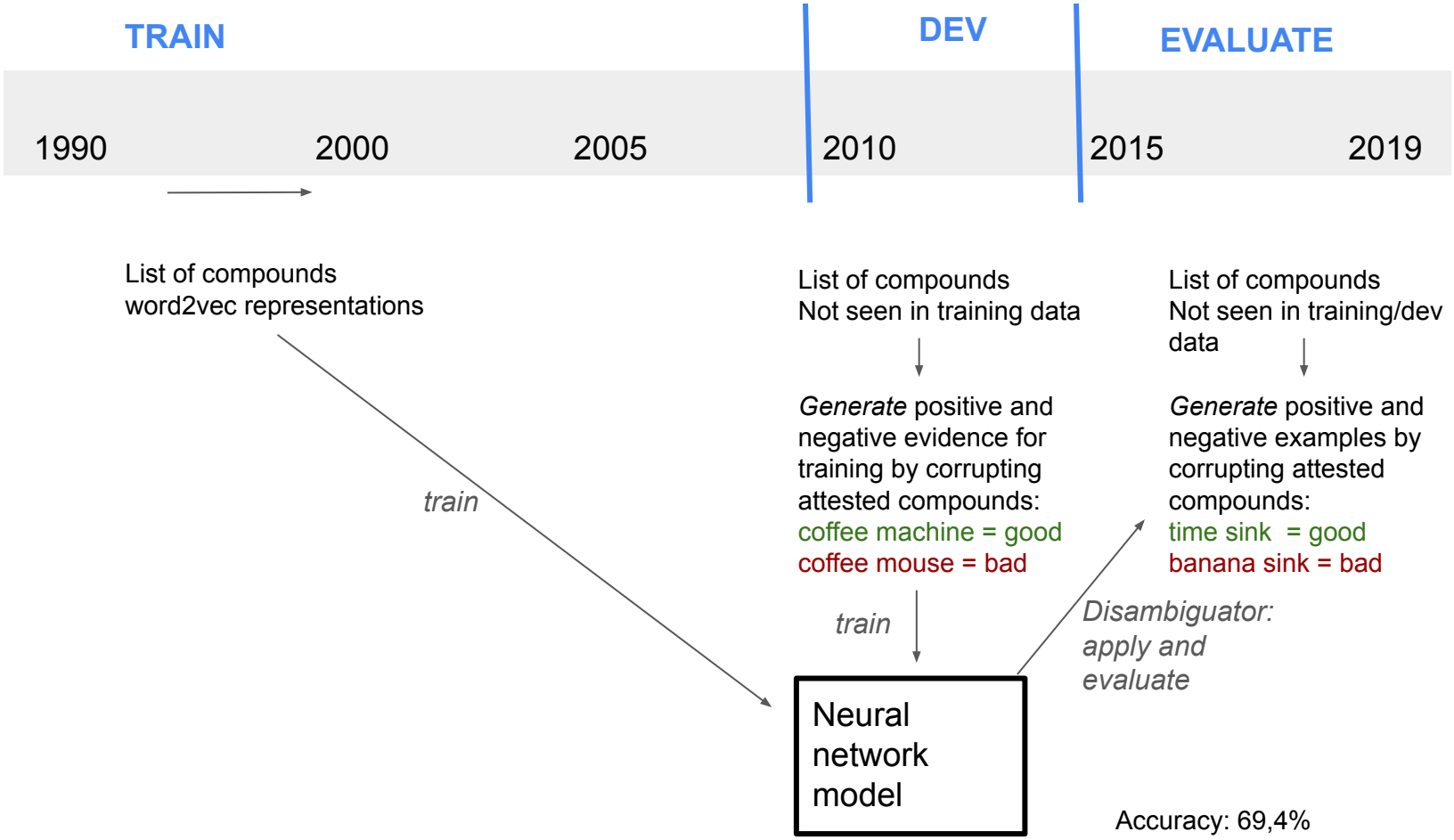
Vector-based
semantic model
of words and
their relations

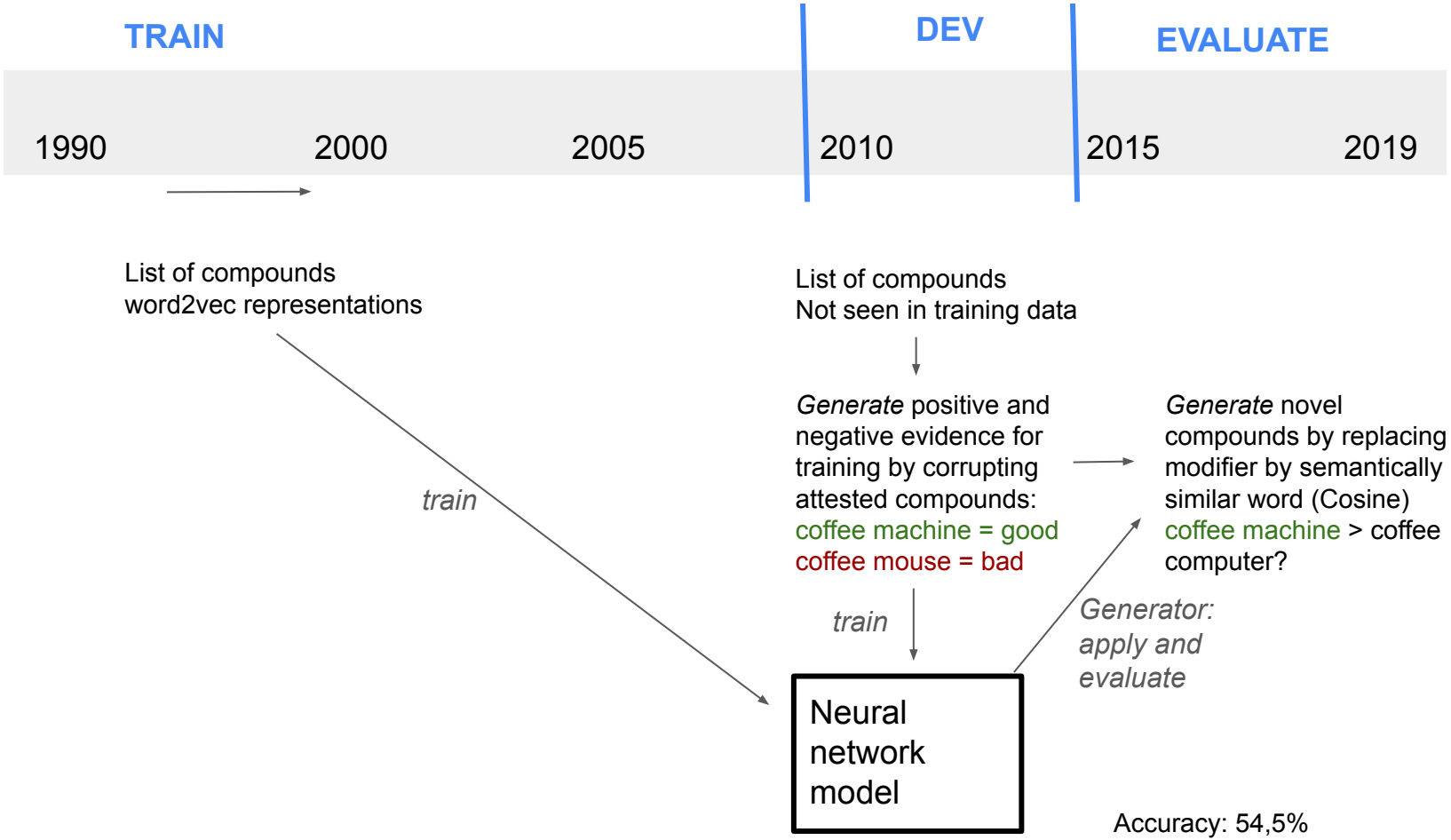


coffee machine = good
coffee mouse = bad

Train model
to learn what are
good combinations

Predict
new ideas

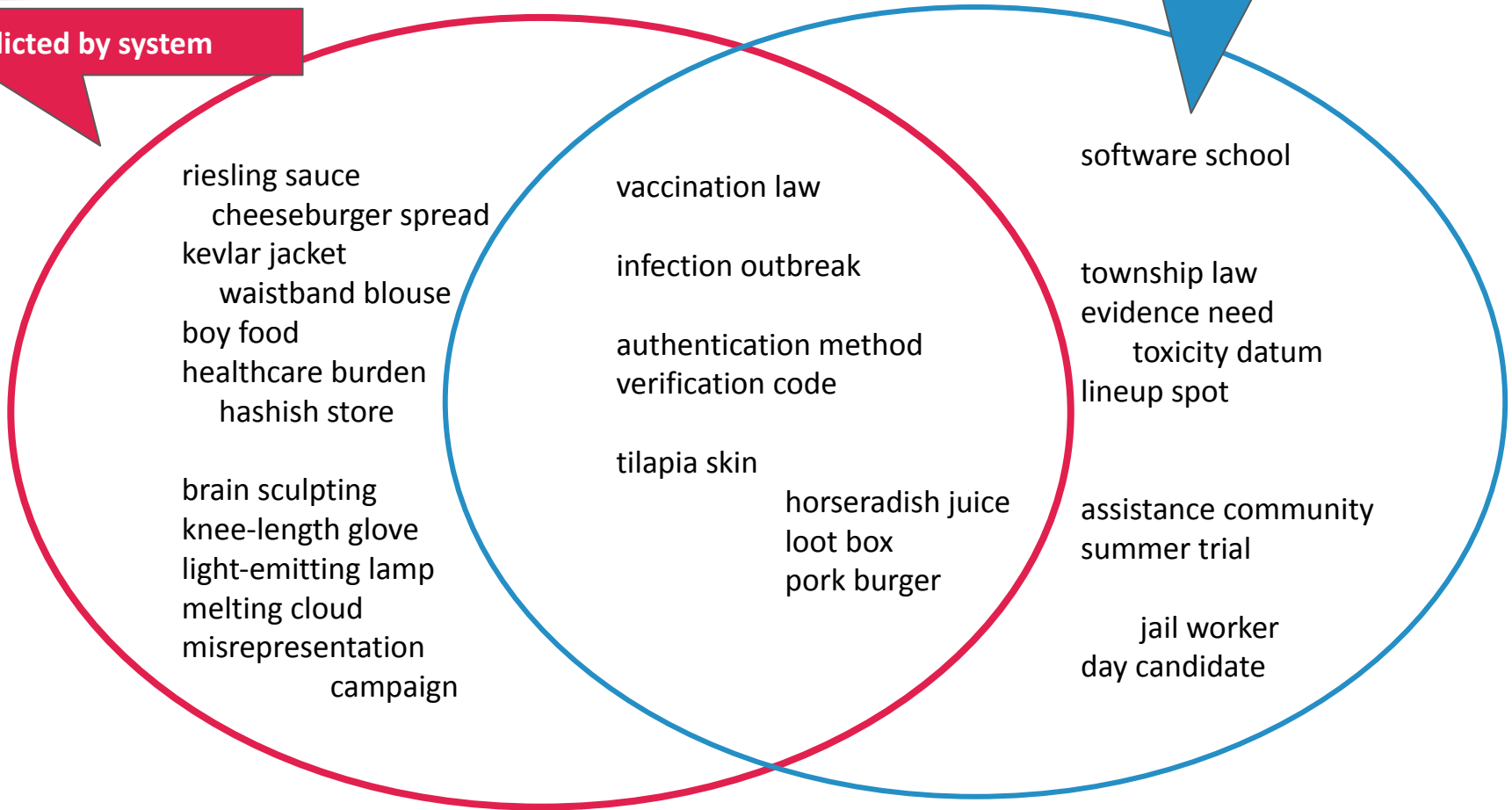

TRAIN
DEV
EVALUATE
1990
2000
2005
2010
2015
2019
List of compounds
word2vec representations
train
List of compounds
Not seen in training data
Generate positive and negative evidence for training by corrupting attested compounds:
coffee machine = good
coffee mouse = bad
train
List of compounds
Not seen in training/dev data
Generate positive and negative examples by corrupting attested compounds:
time sink = good
banana sink = 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)
coffee machine > coffee computer?
Generator: apply and evaluate
Neural network model
Accuracy: 54,5%

System output Generator

Predicted by system

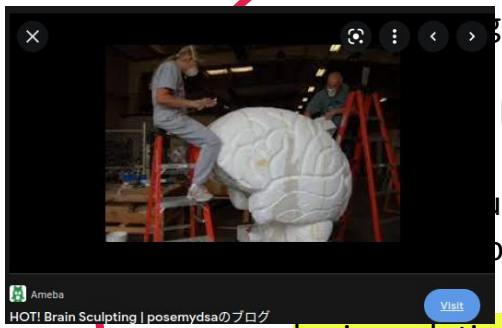
Found in evaluation set 2015-2019



System output Generator

Found in evaluation set 2015-2019

Predicted by system



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

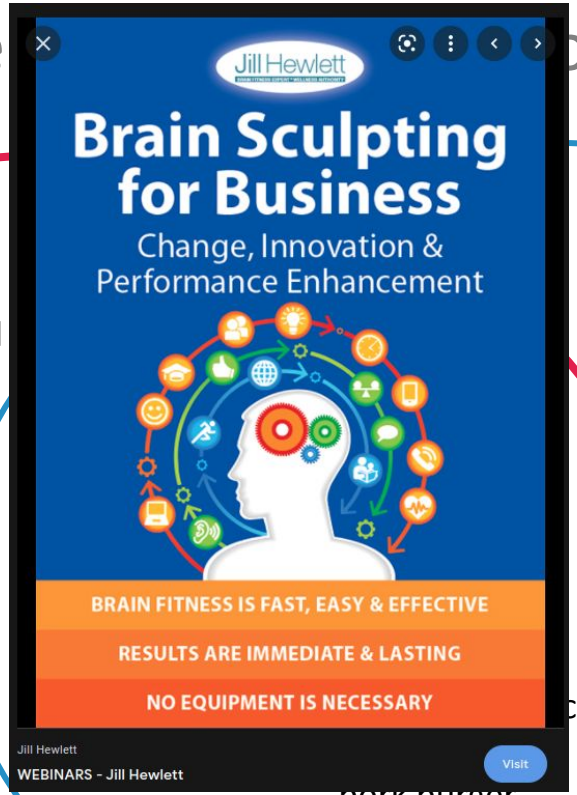
software school
township law
evidence need
toxicity datum
lineup spot
assistance community
summer trial
jail worker
day candidate

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

ce
pork burger

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

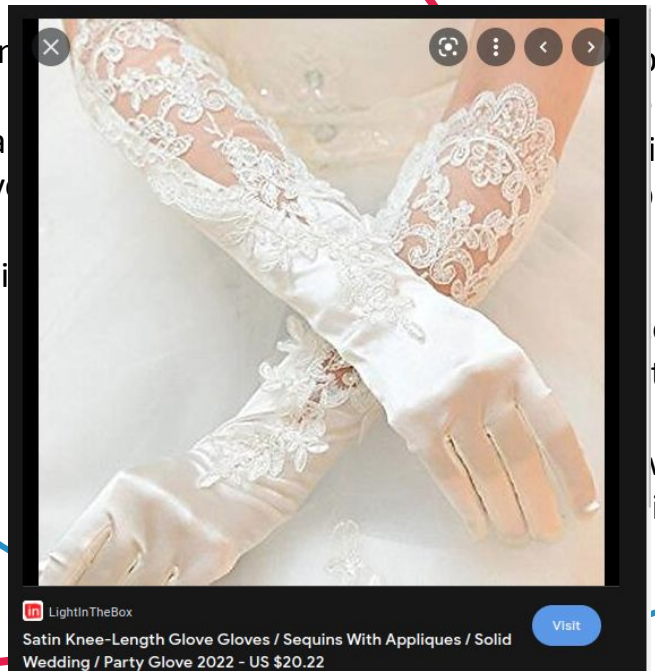
Predicted by system

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



L'évaluation est un défi

- Un certain nombre de composés parmi nos faux positifs semblent être de vrais positifs
- Comment le déterminer automatiquement ?
- Recherches Google (comptes) ?
- Cependant, il est difficile de choisir un seuil approprié pour ce que nous pouvons considérer comme un « bon » composé.

Évaluation est un *challenge*

- Prenez un échantillon aléatoire de 100 mots composés de notre liste de faux positifs.
- Obtenez ensuite le nombre d'occurrences renvoyées par Google lors de la recherche de chaque mot composé.

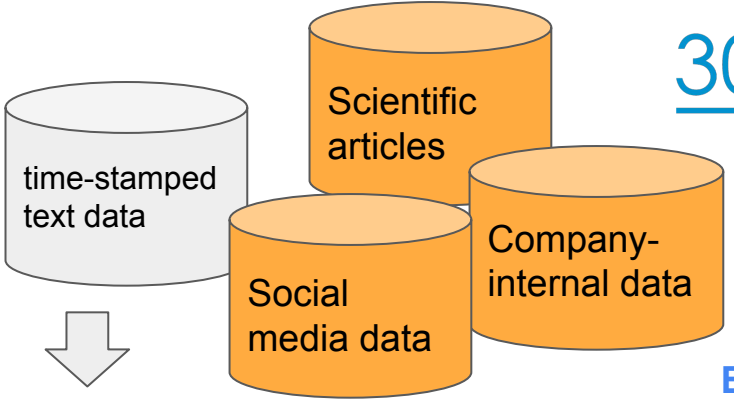
Threshold: minimum 5000 Google hits
Percentage of ‘correct’ compounds among false positives: 66%
Adjusted accuracy if this were true: 84,48%

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

Counted as correct:	Counted as incorrect:
cash counter	mucous defect
porky dog	mistreatment
	complaint
	heron tooth

Counted as correct:	Counted as incorrect:
Snowmobile rental	midmorning train
heel sandal	porky dog

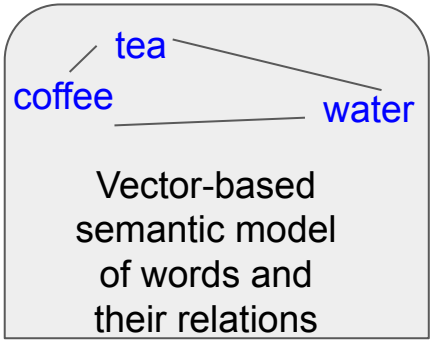
International Create Challenge '21



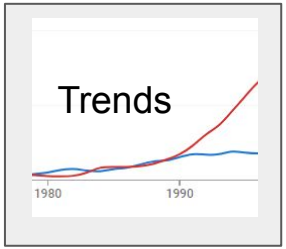
TRAIN

EVALUATE

Time



Vector-based semantic model of words and their relations



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

C-LING : vers des systèmes Créatifs avec modélisation LINGuistique

Projet Fond National Suisse de la Recherche Scientifique (FNS) : 2 doctorants travaillent sur le sujet

Résolution créative de problèmes

Inclut également des modèles interdomaines et interlinguistiques

Capacités créatives des LLMs génératives

- Il est facile de voir un comportement créatif
- Nous (Prof. Antoine Bosselut (EPFL), Soyoung Oh et moi) collaborons avec des chercheurs en sciences cognitives de l'UVA pour tester les capacités créatives de ces systèmes
- Les tests de créativité humaine ne sont pas adaptés pour les systèmes
- La fixation semble être un inconvénient des LLMs

Résumé

- Histoire des modèles de langage naturel
- Quelques menaces et opportunités du NLP
- Exemples de recherche de mon équipe

Quelques idées de collaboration

Education de l'IA (technique, éthique, risques)

Outils qui peuvent aider les enseignants

Quelles opportunités voyez-vous?

Merci de votre attention!