



Quelques bases de Traitement Automatique **des** Langues (TAL) : une décennie de révolutions

Karën Fort

karen.fort@univ-lorraine.fr / <https://members.loria.fr/KFort/>

Quelques sources d'inspiration

- ▶ Cours de Xavier Tannier, ETAL 2023, Marseille
- ▶ [Activité débranchée](#) de Marie Duflot-Kremer sur les réseaux de neurones, elle-même créée à partir de [Brain in a bag](#) de Teaching London Computing
- ▶ Sur les plongements : Comprendre et utiliser les word embeddings, de Bénédicte Pierrejean (CLLE-ERSS)
- ▶ (excellente) Vidéo de Machine Learnia sur les [bases du Deep Learning](#)

Sources

Notions clés (rappels)

Un peu d'histoire

Les plongements statiques

Les plongements contextuels

Rappels : morphologie

Largement inspiré de Xavier Tannier (ETAL 2023)

Flexion

- ▶ Verbale : montrer, montreras...
- ▶ Nominale : cheval, chevaux...
- forme canonique (lemme) vs formes fléchies

Dérivation

- ▶ penser/V + able = pensable
- ▶ in + pensable/A = impensable
- base vs dérivé

Composition

- ▶ appendice + ectomie = appendicectomie
- éléments de formation, mot composé

Rappels : lemmatisation

Largement inspiré de Xavier Tannier (ETAL 2023)

Obtention de la forme canonique (lemme) à partir du mot :

- ▶ Pour un verbe : sa forme à l'infinitif (sans les flexions) montrer, montreras, montraient → montrer
- ▶ Pour un nom, adjectif, article, ... : sa forme au masculin singulier vert, vertes, verts → vert

La lemmatisation demande des ressources et un traitement linguistique

- ▶ En particulier pour les nombreuses exceptions
- ▶ Long et donc difficile à mettre en œuvre pour des grandes collections
- ▶ Dépendant de la langue

Elle n'agrège que des variantes flexionnelles

- ▶ cheval = chevaux
- ▶ cheval ≠ chevalier

Rappels : racinisation (*stemming*)

Largement inspiré de Xavier Tannier (ETAL 2023)

Obtention de la racine, une forme tronquée du mot, commune à toutes les variantes morphologiques

- ▶ Suppression des flexions
- ▶ Suppression des suffixes
- ▶ Ex : cheval, chevaux, chevalier, chevalerie, chevaucher → "cheva" (mais pas "cavalier")

La racinisation est généralement à base de règles

- ▶ Rapide
- ▶ Dépendant de la langue

Elle agrège beaucoup plus que la lemmatisation

- ▶ Vocabulaire plus petit

Sous le mot

Largement inspiré de Xavier Tannier (ETAL 2023)

Subwords

- ▶ n-grammes de caractères
- ▶ Puis agrégation des sous-mots en mots (somme des vecteurs)

Exemple : school

{'sch', 'cho', 'hoo', 'ool', 'scho', 'choo', 'hool', 'schoo', 'chool', 'school'}

WordPieces

- ▶ Un vocabulaire de taille prédéfinie, composé de n-grammes de caractères
- ▶ Vocabulaire choisi pour maximiser la fréquence des n-grammes
- ▶ Possibilité d'un tokenizer multilingue

Exemple (FlauBERT) :

nous uti ##lisons des mo ##deles de re ##presentation contextu ##elle

Sources

Notions clés (rappels)

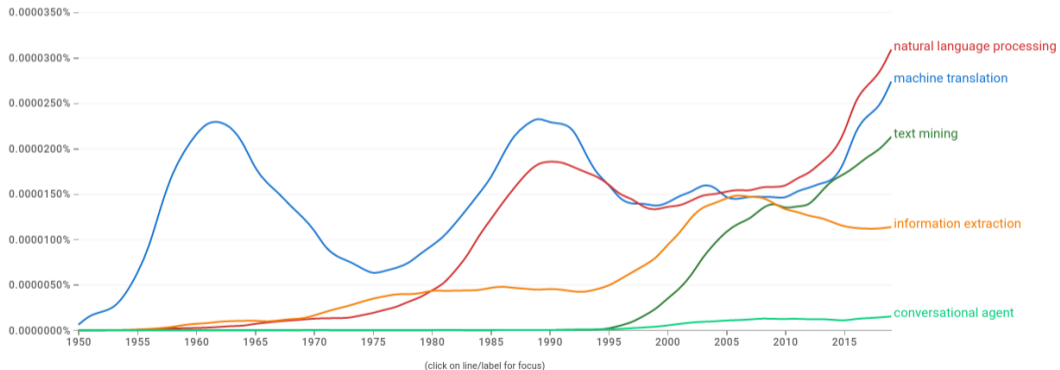
Un peu d'histoire

Les plongements statiques

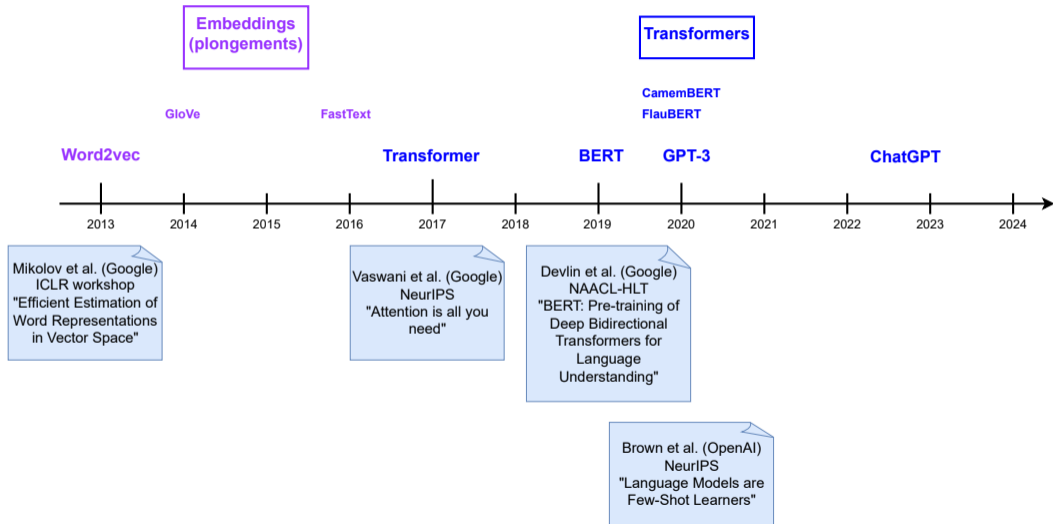
Les plongements contextuels

machine translation,natural language processing,text mining,information extractic

1950 - 2019 English (2019) Case-Insensitive Smoothing



Une décennie révolutionnaire pour le TAL (et l'IA)



Des révolutions qui viennent de loin : le premier modèle de langue

1. Zero-order approximation (symbols independent and equiprobable).

XFOML RXKHRJFFJUJ ZLPWCFWKCYJ FFJEYVKCQSGHYD QPAAMKBZAACIBZL-
HJQD.

2. First-order approximation (symbols independent but with frequencies of English text).

OCRO HLI RGWR NMIELWIS EU LL NBNSEBYA TH EEI ALHENHTTPA OOBTTVA
NAH BRL.

3. Second-order approximation (digram structure as in English).

ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY ACHIN D ILONASIVE TU-
COOWE AT TEASONARE FUSO TIZIN ANDY TOBE SEACE CTISBE.

4. Third-order approximation (trigram structure as in English).

IN NO IST LAT WHEY CRATICT FROURE BIRS GROCID PONDENOME OF DEMONS-
TURES OF THE REPTAGIN IS REGOACTIONA OF CRE.

5. First-order word approximation. Rather than continue with tetragram, \dots , n -gram structure it is easier and better to jump at this point to word units. Here words are chosen independently but with their appropriate frequencies.

REPRESENTING AND SPEEDILY IS AN GOOD APT OR COME CAN DIFFERENT NAT-
URAL HERE HE THE A IN CAME THE TO OF TO EXPERT GRAY COME TO FURNISHES
THE LINE MESSAGE HAD BE THESE.

6. Second-order word approximation. The word transition probabilities are correct but no further structure is included.

THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHAR-
ACTER OF THIS POINT IS THEREFORE ANOTHER METHOD FOR THE LETTERS THAT
THE TIME OF WHO EVER TOLD THE PROBLEM FOR AN UNEXPECTED.

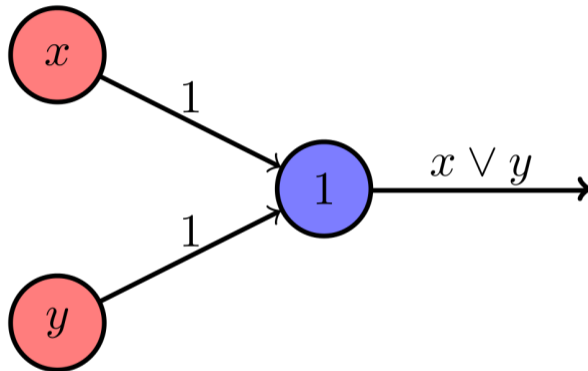
Des révolutions qui viennent de loin : l'hypothèse distributionnelle

with which they are actually found. More than that: if we consider words or morphemes A and B to be more different in meaning than A and C, then we will often find that the distributions of A and B are more different than the distributions of A and C. In other words, difference of meaning correlates with difference of distribution.

<https://www.tandfonline.com/doi/pdf/10.1080/00437956.1954.11659520>

[Harris, 1954]

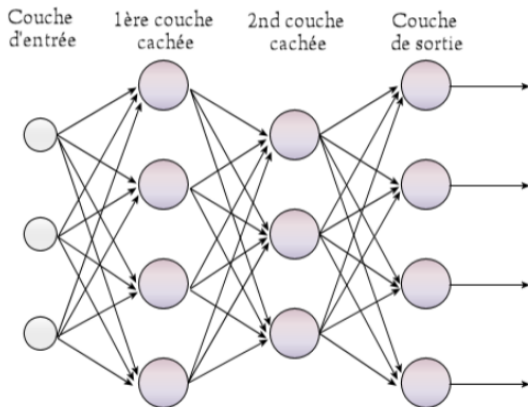
Des révolutions qui viennent de loin : le perceptron



MartinThoma <https://fr.wikipedia.org/wiki/Perceptron>

[Rosenblatt, 1958]

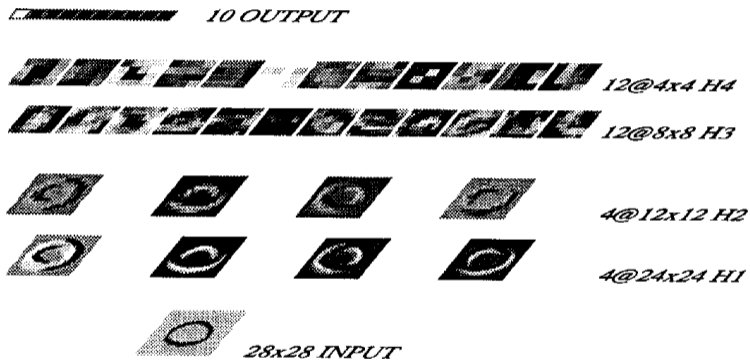
Des révolutions qui viennent de loin : le perceptron multicouches et l'algorithme de *backpropagation*



[Rumelhart et al., 1986]

Des révolutions qui viennent de loin : la première application réelle

la reconnaissance de codes postaux écrits manuellement



[LeCun et al., 1990]

Des révolutions qui viennent de loin : le premier modèle de langue neuronal

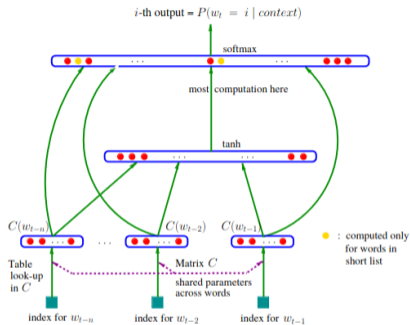


Figure 1: “Direct Architecture”: $f(i, w_{t-1}, \dots, w_{t-n}) = g(i, C(w_{t-1}), \dots, C(w_{t-n}))$
where g is the neural network and $C(i)$ is the i -th word feature vector.

[Bengio et al., 2000]

Vidéo

<https://www.youtube.com/watch?v=XUFLq6dKQok>

Sources

Notions clés (rappels)

Un peu d'histoire

Les plongements statiques

Les plongements contextuels

Plongements lexicaux (*word embeddings*)

Largement inspiré de Xavier Tannier (ETAL 2023)

- ▶ **Intuition 1** : Chaque mot d'une langue est associé à une composition de facteurs cachés (souvent inintelligibles)
 - Ex : chat = 10 (animal) + 5 (doux) - 10 (loyal)
- ▶ **Intuition 2** : deux mots proches dans l'espace vectoriel = deux mots qui partagent souvent des contextes similaires (hypothèse distributionnelle)
 - Ex : le ... griffe; ... est un félin

$$\text{occurrence}(\text{chat}) \sim \text{occurrence}(\text{tigre})$$

$$W_{\text{chat}} \cdot W_{\text{contexte}} \sim W_{\text{tigre}} \cdot W_{\text{contexte}}$$

$$W_{\text{chat}} \sim W_{\text{tigre}}$$

(©Perceval Wajsbürt)

Visualiser les plongements lexicaux

<http://projector.tensorflow.org/>

Démo sur Google Collab

Sources

Notions clés (rappels)

Un peu d'histoire

Les plongements statiques

Les plongements contextuels

Bibliographie

Statique vs contextuel

Représentation statique : un token = un vecteur

- ▶ On manipule une « matrice d'embeddings » ($N \times d$)
- ▶ Le vecteur du token est le même à chacune de ses occurrences dans le corpus

vs

Représentation contextuelle : calcul du vecteur en contexte

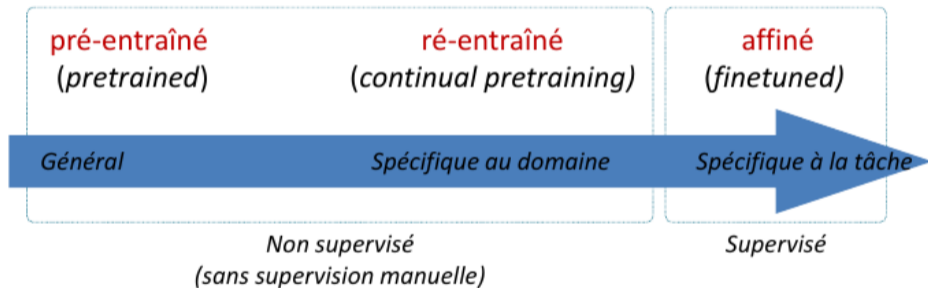
- ▶ Le calcul de la représentation est intégré dans le modèle
- ▶ Les mots précédents et suivants agissent sur la représentation (en général grâce à un mécanisme d'attention...)

Modèles de langues : type d'application

Modèles sans supervision manuelle, avec deux principaux types de **pré-entraînement** :

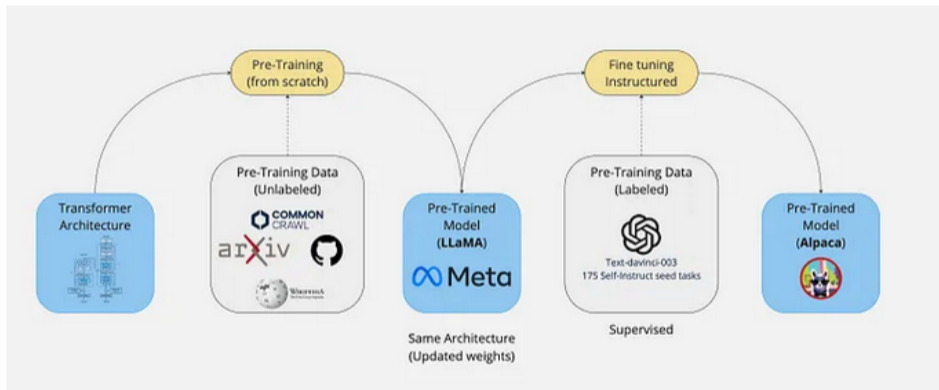
- ▶ **Prédire le mot suivant** : modèles autorégressifs (par ex. GPT)
- ▶ **Prédire des mots masqués** dans une séquence : modèles de langue masqués ou MLM (par ex. BERT)

Pré-entraînement vs affinage vs pré-entraînement continu



Pré-entraînement vs affinage : adaptation à une tâche

Alpaca est l'agent conversationnel dérivé de LLaMa (spécialisation)



Example of fine-tuning a LLaMA-based model (Image created by the author)

<https://medium.com/@eordaxd/fine-tuning-vs-pre-training-651d05186faf>

Autre utilisation de l'affinage : adaptation à une langue

Vigogne est la version française de LLaMa



Bofeng Huang • 2nd

NLP Research Engineer @Zaion | CentraleSupélec

7mo • Edited • 🌐

+ Follow ...

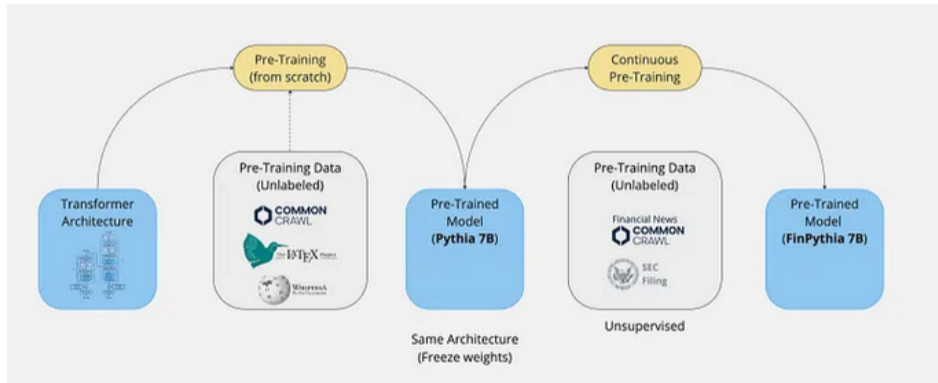
Il y a 2 jours, Meta a publié les modèles Llama-2 🌟, qui bénéficient d'un pré-entraînement sur 2T de tokens, avec une licence plus conviviale pour une utilisation commerciale, et présentent des avancées notables en RLHF.

Malgré leurs performances impressionnantes en anglais, ces modèles ne s'adaptent pas aussi bien aux autres langues, telles que le français. C'est pourquoi nous avons pris l'initiative de les fine-tuner pour qu'ils puissent mieux comprendre et suivre les instructions en français. Le premier modèle ayant terminé ce processus de sft sur Llama-2-7B a été nommé Vigogne-2-7B-Instruct, et il devient le nouveau membre de la famille Vigogne 🐶.

<https://www.linkedin.com/feed/update/urn:li:activity:7087785080881885184/>

Pré-entraînement vs pré-entraînement continu : adaptation à un domaine

FinPythia est l'adaptation de Pythia à la finance (*transfer learning*)



Example of further pre-train a Pythia based model (Image created by the author)

<https://medium.com/@eordaxd/fine-tuning-vs-pre-training-651d05186faf>

L'exemple de LLaMa 2 - Chat : entraînement

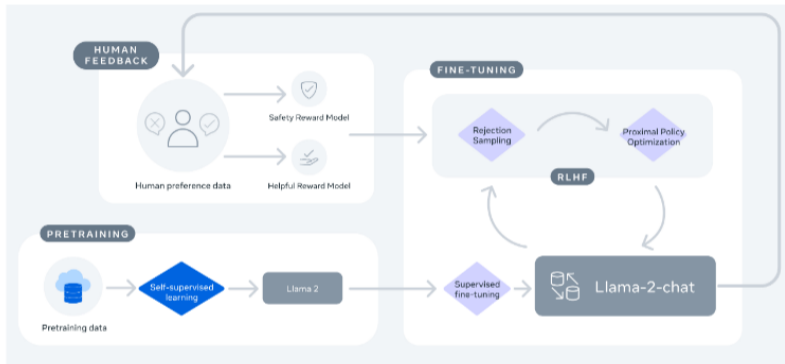


Figure 4: Training of LLAMA 2-CHAT: This process begins with the **pretraining** of LLAMA 2 using publicly available online sources. Following this, we create an initial version of LLAMA 2-CHAT through the application of **supervised fine-tuning**. Subsequently, the model is iteratively refined using Reinforcement Learning with Human Feedback (**RLHF**) methodologies, specifically through rejection sampling and Proximal Policy Optimization (PPO). Throughout the RLHF stage, the accumulation of **iterative reward modeling data** in parallel with model enhancements is crucial to ensure the reward models remain within distribution.

<https://arxiv.org/pdf/2307.09288.pdf>

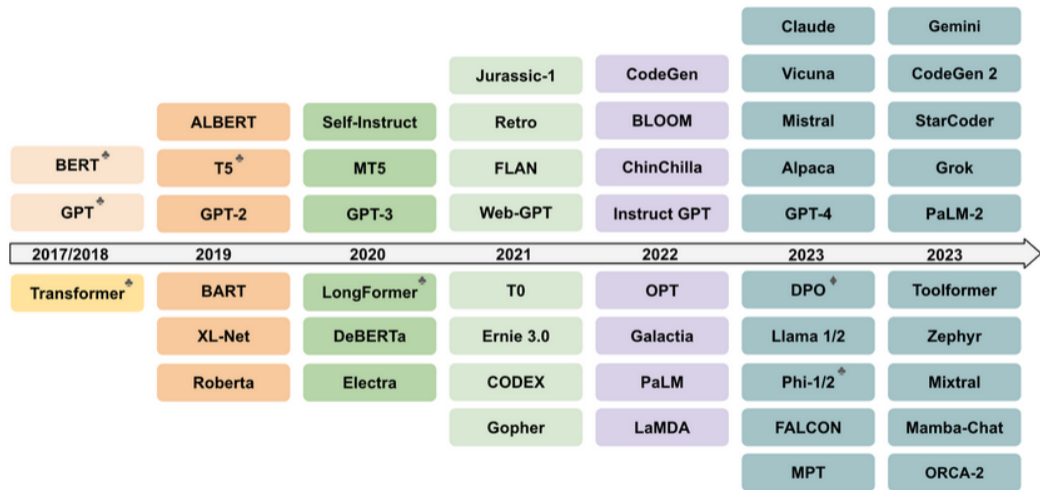
L'exemple de LLaMa 2 - Chat : langues

| Language | Percent | Language | Percent |
|----------|---------|----------|---------|
| en | 89.70% | uk | 0.07% |
| unknown | 8.38% | ko | 0.06% |
| de | 0.17% | ca | 0.04% |
| fr | 0.16% | sr | 0.04% |
| sv | 0.15% | id | 0.03% |
| zh | 0.13% | cs | 0.03% |
| es | 0.13% | fi | 0.03% |
| ru | 0.13% | hu | 0.03% |
| nl | 0.12% | no | 0.03% |
| it | 0.11% | ro | 0.03% |
| ja | 0.10% | bg | 0.02% |
| pl | 0.09% | da | 0.02% |
| pt | 0.09% | sl | 0.01% |
| vi | 0.08% | hr | 0.01% |





Table 10: Language distribution in pretraining data with percentage $\geq 0.005\%$. Most data is in English, meaning that LLAMA 2 will perform best for English-language use cases. The large unknown category is partially made up of programming code data.





<https://arxiv.org/pdf/2307.09288.pdf>

L'explosion des LLM



<https://arxiv.org/html/2402.06196v1>

-  Bannour, N., Ghannay, S., Névéol, A., and Ligozat, A.-L. (2021).
Evaluating the carbon footprint of NLP methods : a survey and analysis of existing tools.
In EMNLP, Workshop SustaiNLP, Punta Cana, Dominican Republic.
-  Bengio, Y., Ducharme, R., and Vincent, P. (2000).
A neural probabilistic language model.
In Leen, T., Dietterich, T., and Tresp, V., editors, Advances in Neural Information Processing Systems, volume 13. MIT Press.
-  Harris, Z. (1954).
Distributional structure.
Word, 10(23) :146–162.
-  Hovy, D. and Prabhumoye, S. (2021).
Five sources of bias in natural language processing.
Language and Linguistics Compass, 15(8) :e12432.

-  Kirk, H. R., Jun, Y., Iqbal, H., Benussi, E., Volpin, F., Dreyer, F. A., Shtedritski, A., and Asano, Y. M. (2021).
Bias out-of-the-box : An empirical analysis of intersectional occupational biases in popular generative language models.
In [Neural Information Processing Systems](#).
-  LeCun, Y., Matan, O., Boser, B. E., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. E., Jacket, L. D., and Baird, H. S. (1990).
Handwritten zip code recognition with multilayer networks.
volume ii, pages 35–40 vol.2.
-  Rosenblatt, F. (1958).
The perceptron : A probabilistic model for information storage and organization in the brain.
[Psychological Review](#), 65(6) :386–408.
-  Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986).
[Learning internal representations by error propagation](#), pages 318–362.
MIT Press, Cambridge, MA, USA.

-  Shannon, C. E. (1948).
A mathematical theory of communication.
The Bell System Technical Journal, 27 :379–423.
-  Strubell, E., Ganesh, A., and McCallum, A. (2019).
Energy and policy considerations for deep learning in NLP.
In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.
-  Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. (2017).
Men also like shopping : Reducing gender bias amplification using corpus-level constraints.
In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2979–2989, Copenhagen, Denmark. Association for Computational Linguistics.