



## Deep Learning et NLP

### Applications à l'analyse de sentiments

Damien Douteaux

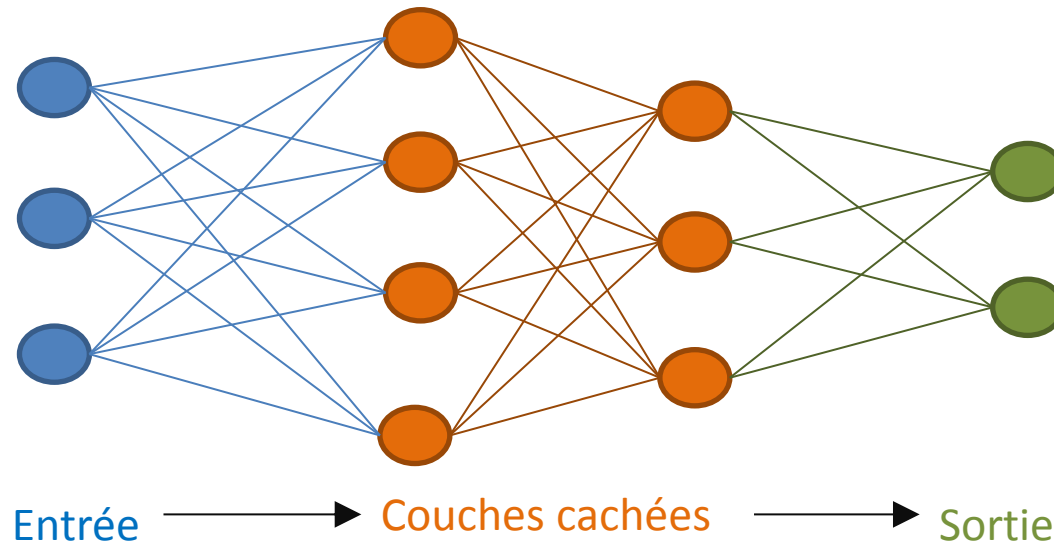
Vincent Hocquemiller

Louis Redonnet

- Contexte et objectifs
- Résultats obtenus (livrables)
- Bilan vis-à-vis des objectifs
- Conclusion

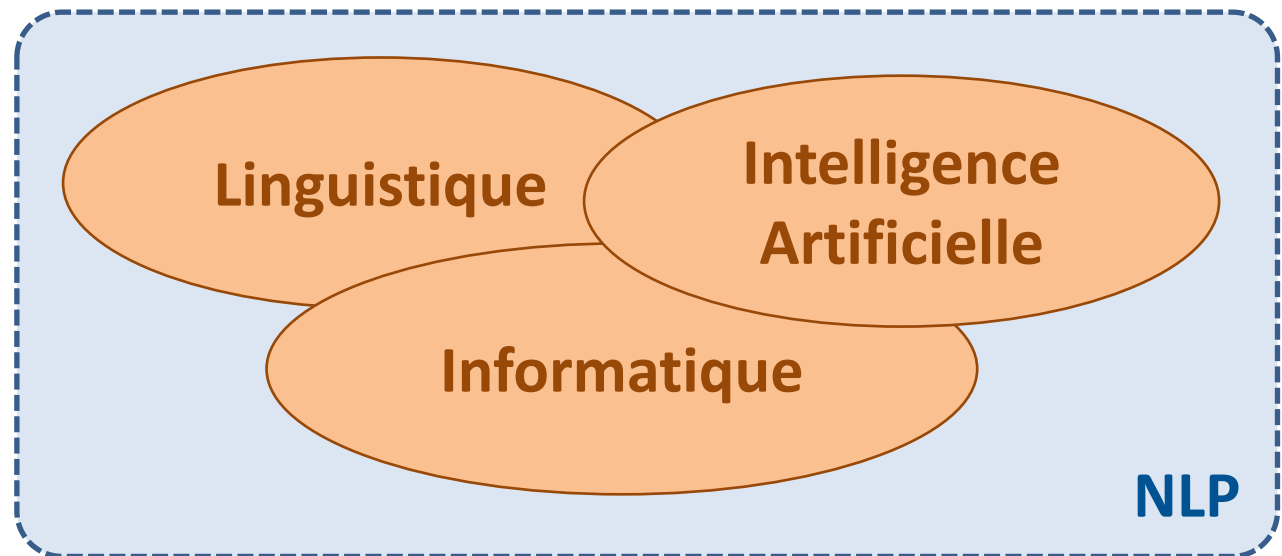
**Deep Learning** et traitement de langues naturelles et Text Mining.

- Méthode récente (1980 – 2000).
- Couches de neurones (unités de traitements).



Deep Learning et **traitement de langues naturelles** et Text Mining.

Le traitement (automatique) des langues naturelles est l'exploitation du langage humain par les outils informatiques.



- Fournir un état de l'art précis sur le contexte.
- Constituer une base de données de volume adapté au Deep Learning.
- Implémentation d'un réseau de neurones sur un cas dégagé par l'état de l'art.

## RECONNAISSANCE D'AUTEUR

---

- Déterminer l'auteur d'un texte.
- Utilisation pour la détection de plagiat/similarité.

## TRADUCTION

---

- Traduction de texte à l'échelle d'une phrase.
- Idée d'essai d'un même réseau pour plusieurs langues.

## INFÉRENCE

---

- Étude de relations logiques entre les phrases d'un corpus.
- Recherche de contradictions ou préservation de la logique.

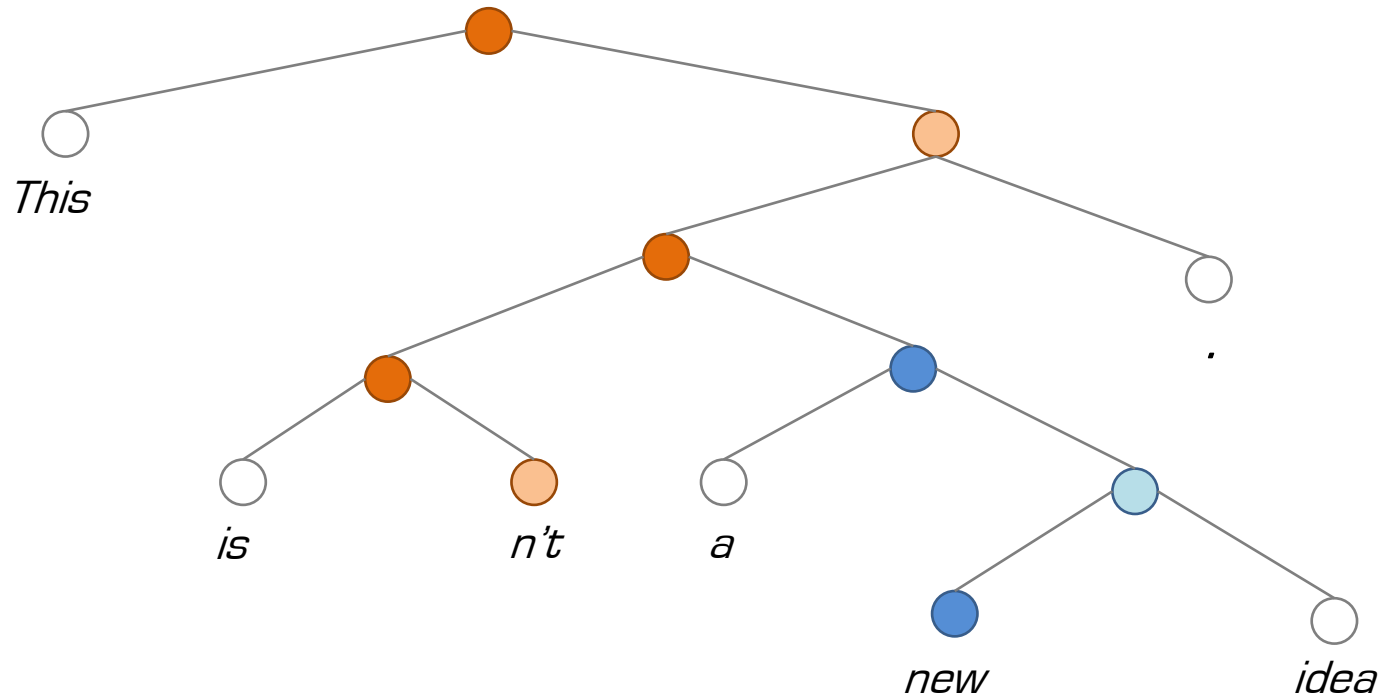
## QUESTIONS-RÉPONSES

---





- Formuler des réponses automatiques à des questions simples.
- Utilisation pour des *chatbox* en ligne.

## ANALYSE DE SENTIMENTS

- Classifier les phrases par sentiment ou essayer de les prédire.
- Différentes structures de données et approches possibles.
- Utilisation en classification de mail, étude d'opinions,...

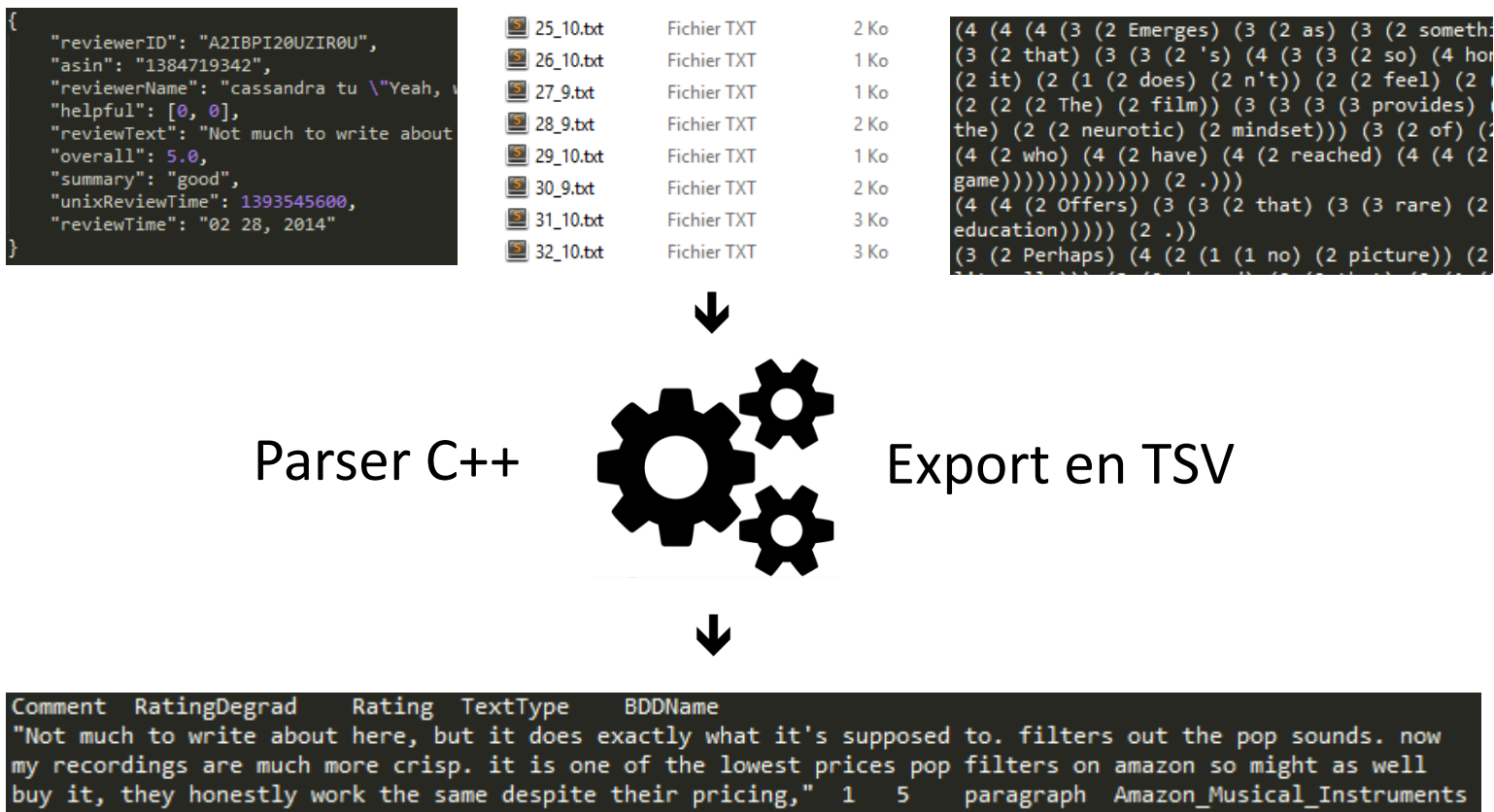


➔ **Sujet retenu pour les essais de la suite du projet.**

	Nom	Quantité de données	Origine
	Large Movie Review Dataset	25000 x 2	Stanford
	Rottent Tomatoes Dataset	215 000	Kaggle
	Twitter Sentiment Corpus	5500	Niek Sanders
	Twitter Sentiment Analysis Corpus	1 578 627	?
	Sentiment Analyses Dataset	9645	Stanford
	UMICH S1650	40 000	Kaggle
	Amazon reviews	> 6 millions <sup>1</sup>	Julian Mc Auley

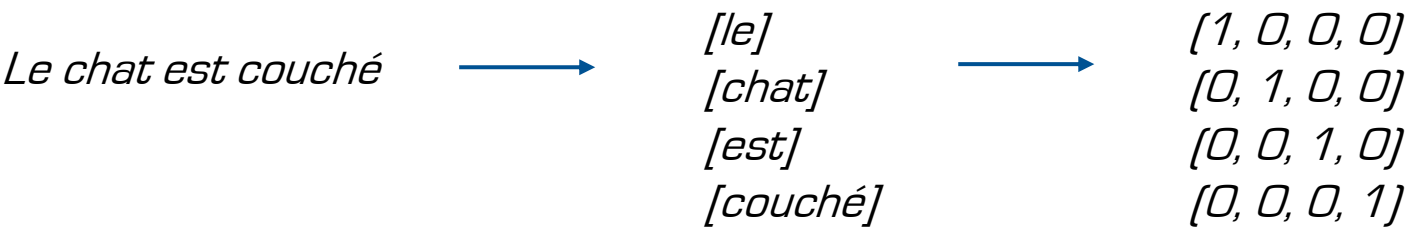
<sup>1</sup> : la base complète (sur demande) fait 142,8 millions de critiques...(20 Gb).



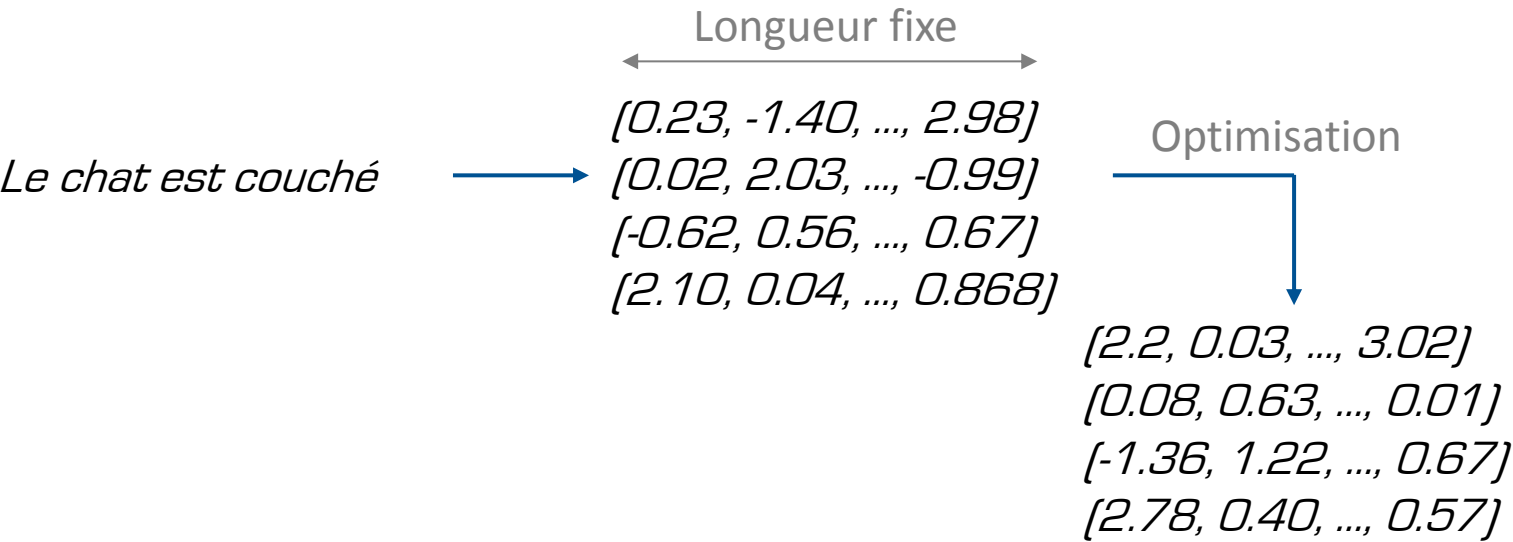


- Format uniforme entre les bases.
- Réduction de la taille.
- Pré-calcul sur la répartition des données.

- Le Deep Learning impose une représentation **Sparse**, **Creuse** ou **Dense**.
- Représentation creuse :

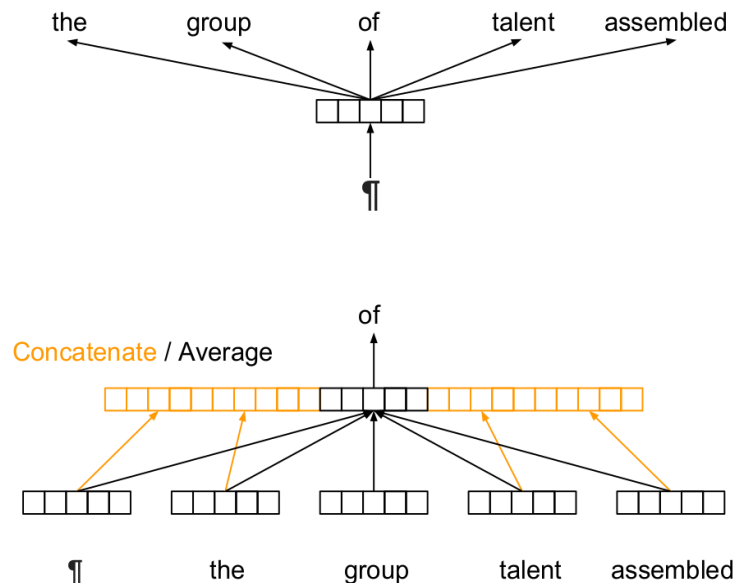


- Représentation dense :



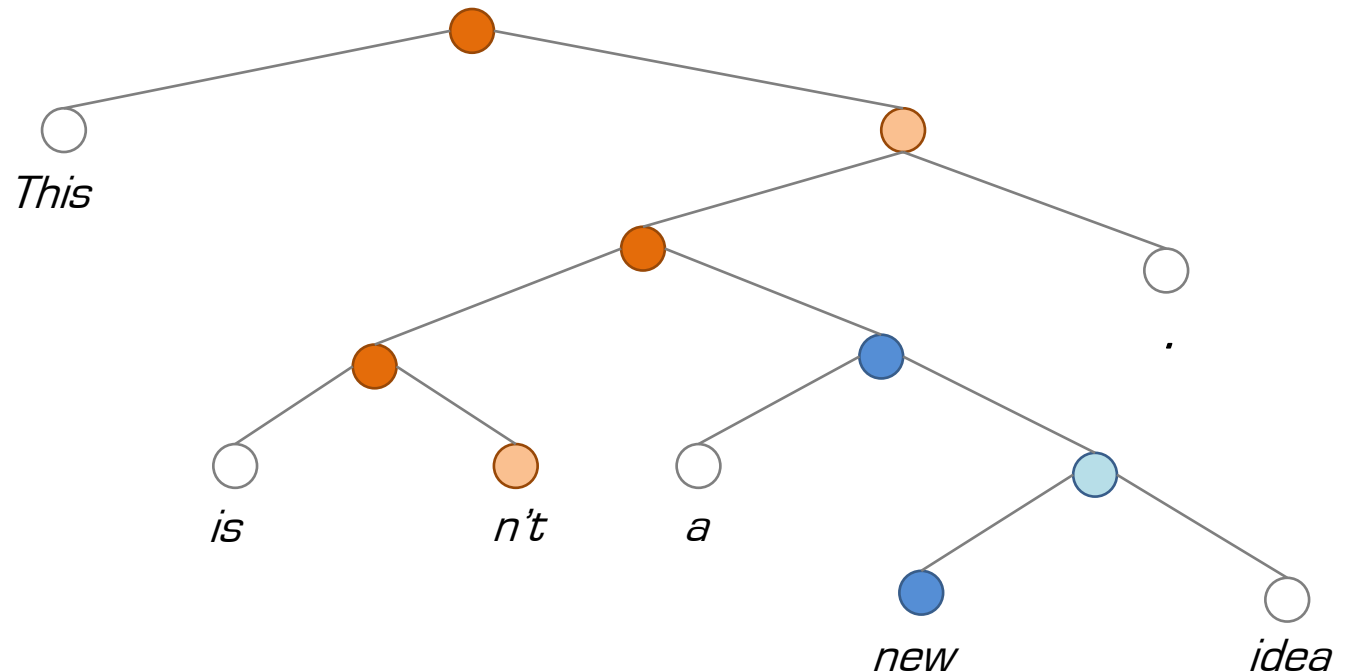
Trois façons de représenter une phrase :

- Un vecteur à apprendre.
- Un arbre descripteur.
- Une suite **ordonnées** de mots (LSTM<sup>1</sup>).



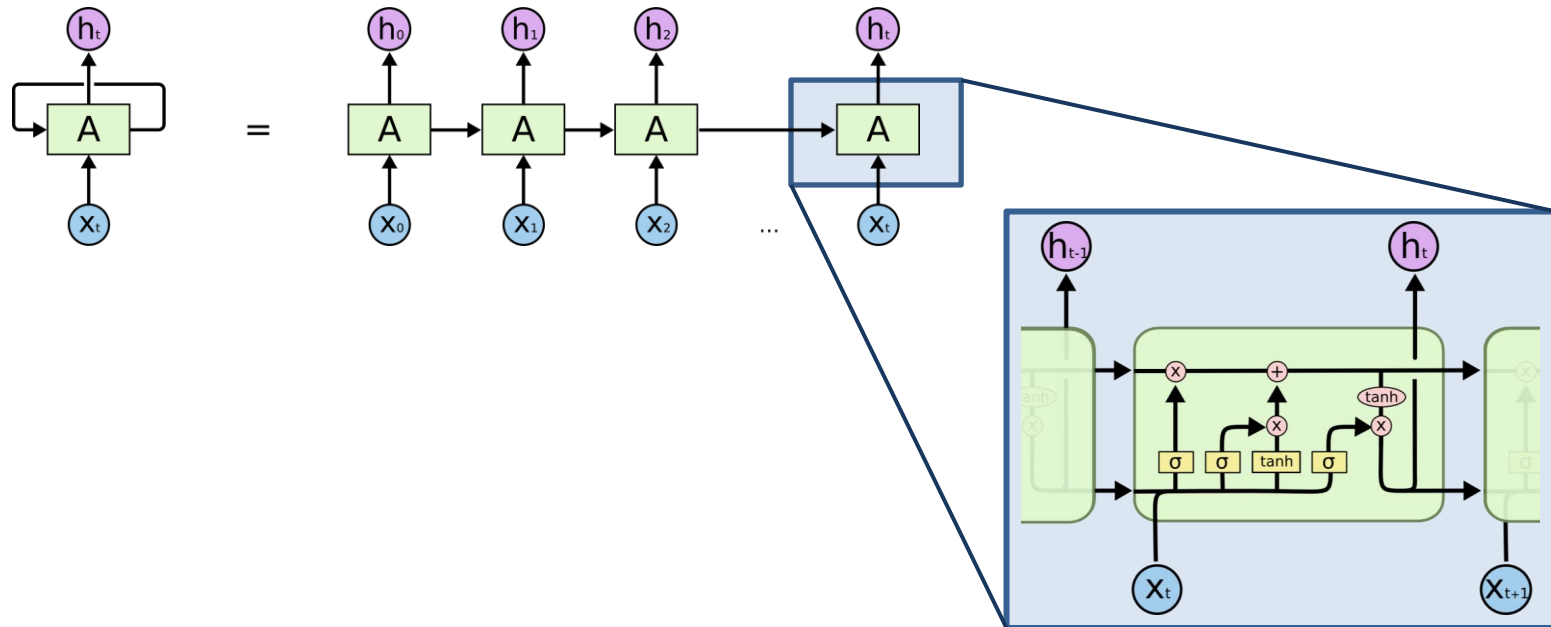
Trois façons de représenter une phrase :

- Un vecteur à apprendre.
- Un arbre descripteur
- Une suite **ordonnées** de mots (LSTM<sup>1</sup>)

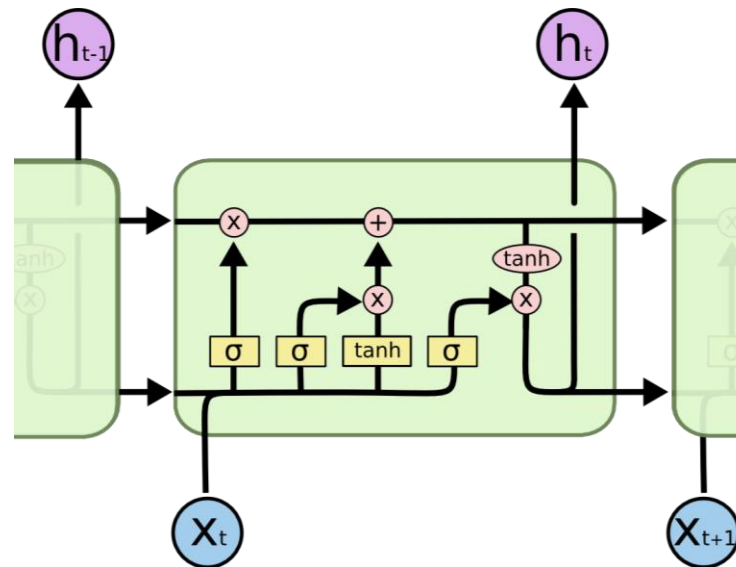
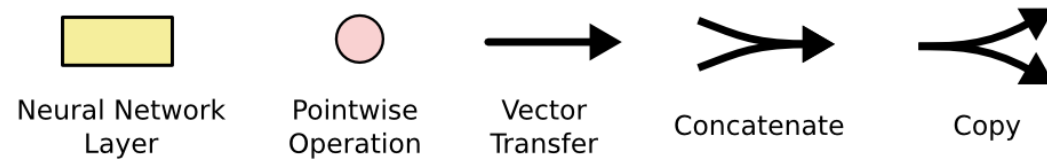


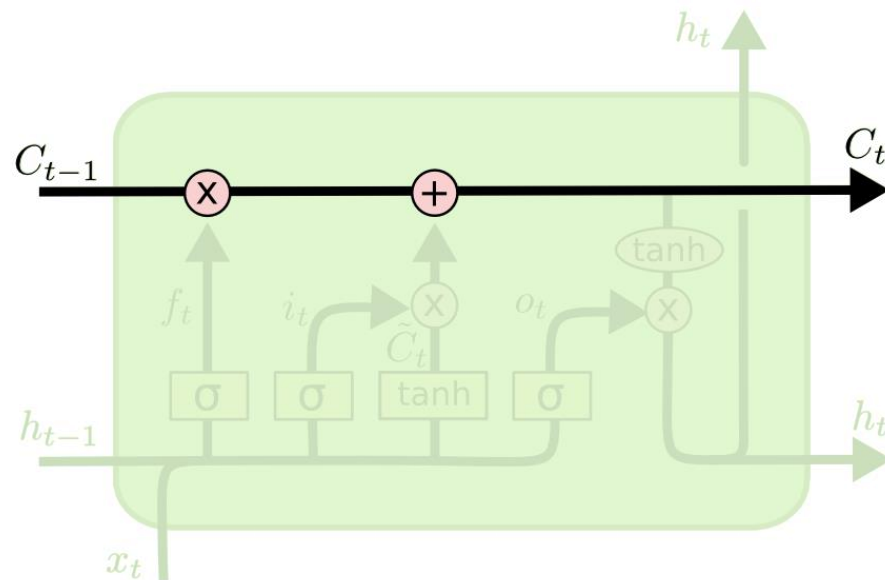
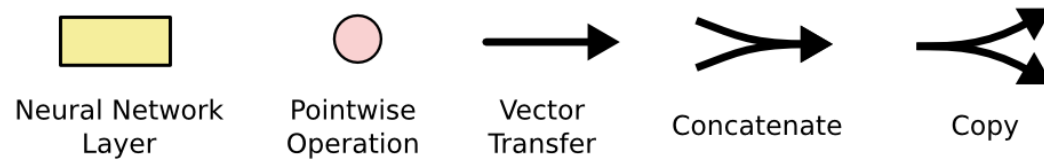
Trois façons de représenter une phrase :

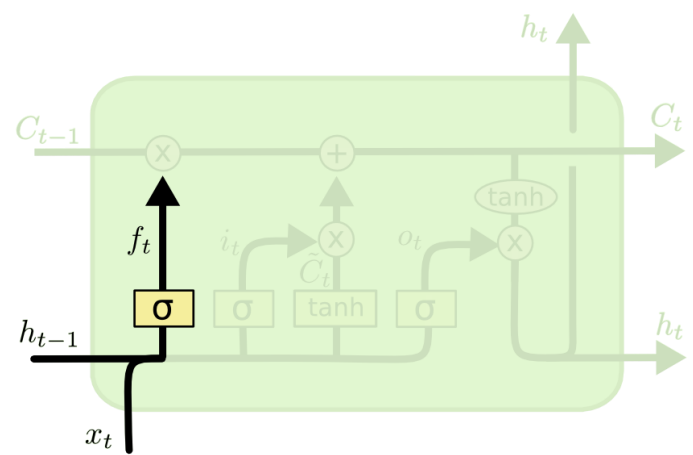
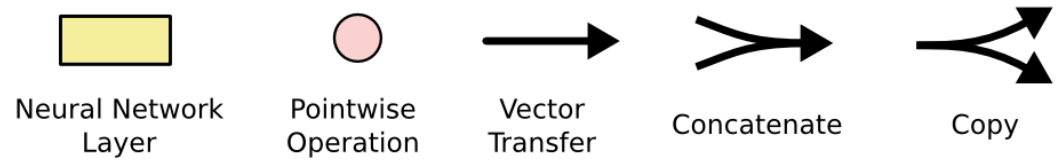
- Un vecteur à apprendre.
- Un arbre descripteur.
- Une suite **ordonnées** de mots (LSTM<sup>1</sup>).



<sup>1</sup> LSTM : Long Short-Term Memory

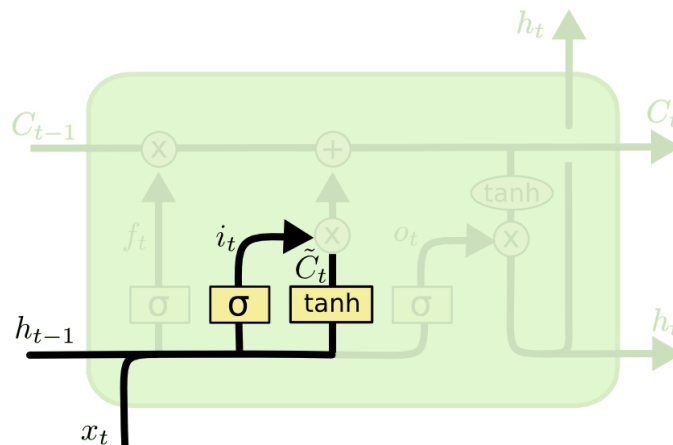
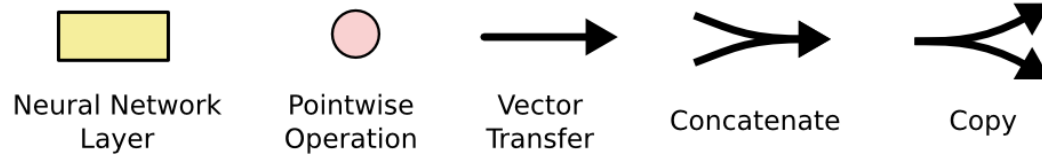




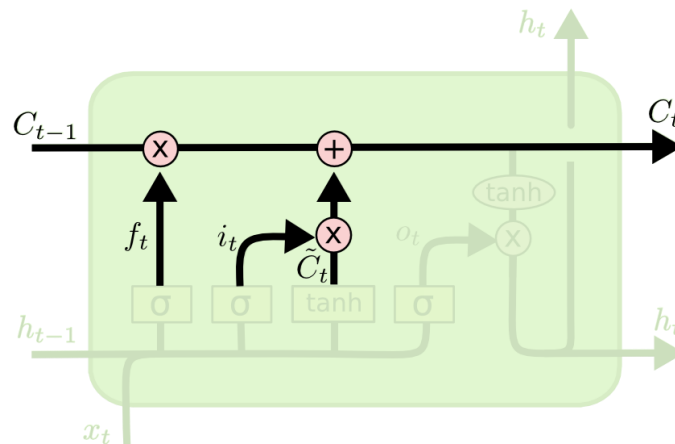
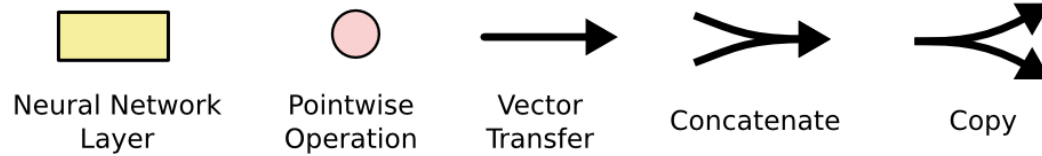


$$f_t = \sigma (W_f \cdot [h_{t-1}, x_t] + b_f)$$

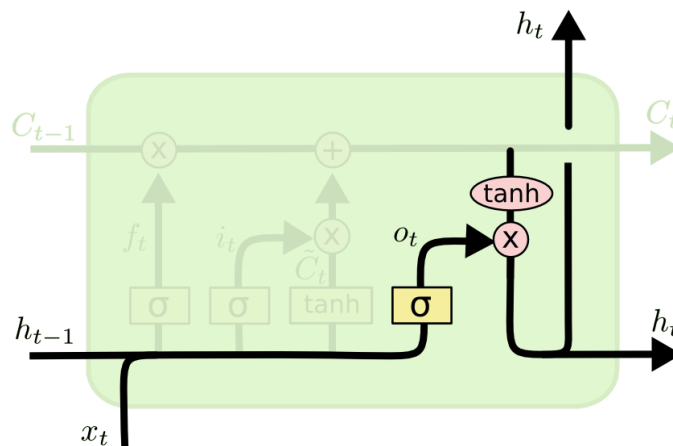
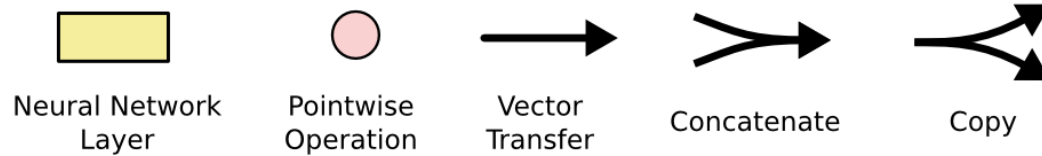




$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$$



$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$



$$o_t = \sigma (W_o [h_{t-1}, x_t] + b_o)$$

$$h_t = o_t * \tanh (C_t)$$

# POURQUOI LES LSTM?

Contexte et  
objectifs

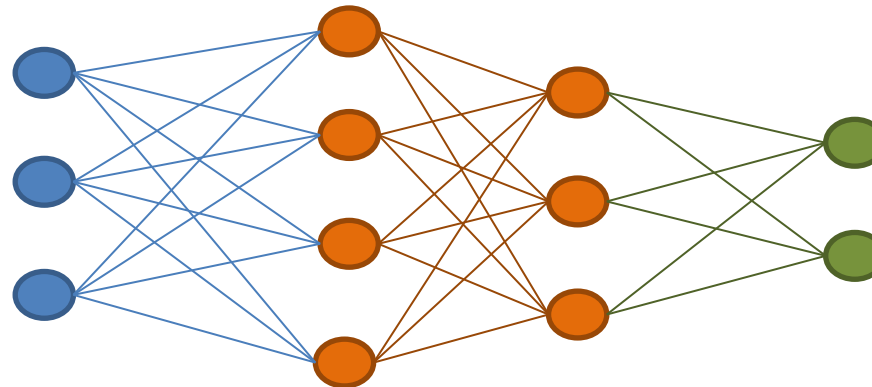
Résultats

Bilan

Conclusion

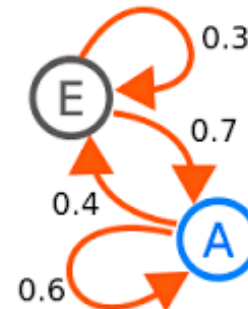
## → Utilisation de données séquentielles.

- Pourquoi ne pas choisir les réseaux classiques (*feed forward*)?

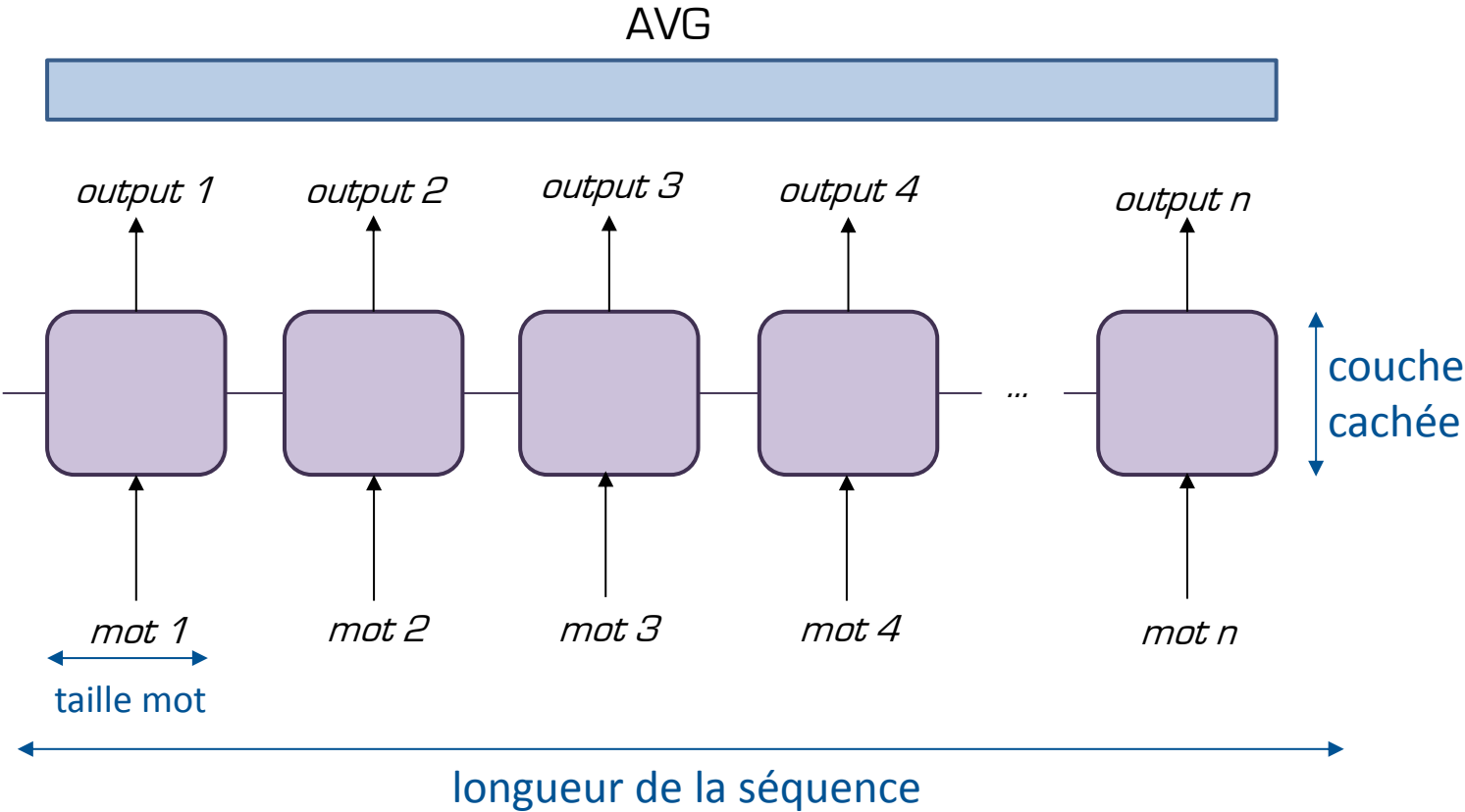


$$\begin{array}{c}
 z \\
 \uparrow \frac{\partial z}{\partial y} \\
 y \\
 \downarrow \frac{\partial y}{\partial x} \\
 x
 \end{array}
 \quad
 \begin{array}{l}
 \Delta z = \frac{\partial z}{\partial y} \Delta y \\
 \Delta y = \frac{\partial y}{\partial x} \Delta x \\
 \Delta z = \frac{\partial z}{\partial y} \frac{\partial y}{\partial x} \Delta x \\
 \frac{\partial z}{\partial x} = \frac{\partial z}{\partial y} \frac{\partial y}{\partial x}
 \end{array}$$

- Pourquoi pas les chaînes de Markov ?



# ARCHITECTURE DU RÉSEAU



## PROBLÈME « JOUET »



- ✓ État de l'art complet sur les applications du Deep Learning au NLP.
- ✓ Récupération d'une base de données adaptée au Deep Learning et prétraitement appliqué.
- ✓ Un *Toy Problem* qui fonctionne, et montre la validité des LSTM.
- ✗ Un début d'application sur nos données.

**➔ Les objectifs initiaux ont été réalisés.**

## ENSEIGNEMENTS

---

- Une occasion d'utiliser les librairies de Deep Learning de Python.
- Un projet enrichissant vis-à-vis des applications abordées.

## PERSPECTIVES

---

- Ne pas s'enfermer que dans *tensorflow* et tester d'autres solutions.
- Pousser plus loin l'application sur les BDD traitées.



**Merci pour votre attention**  
**Et place aux questions!**



[https://github.com/DDouteaux/Projet\\_option\\_info](https://github.com/DDouteaux/Projet_option_info)