Extraction d'Information (Reconnaissance d’entités nommées)

Master DAC, Sorbonne Université

Xavier Tannier
xavier.tannier@sorbonne-universite.fr

Entités nommées

- Entités nommées :
  - Unités lexicales particulières
  - Ex : noms de personnes, noms d'organisation, noms de lieux... dates, unités monétaires, pourcentages...

- Reconnaissance des entités nommées :
  - Identifier ces unités dans un texte
  - Les catégoriser
  - Éventuellement, les normaliser (entity *linking*)

Entités nommées

Identification

Le joueur de tennis américain John McEnroe a déclaré samedi sur ESPN que Gaël Monfils n'était pas assez professionnel. « Monfils aurait déjà dû gagner 4 ou 5 majeurs », a-t-il précisé.

Entités nommées

Catégorisation

Le joueur de tennis américain *personne* John McEnroe a déclaré *date* samedi sur *organisation* ESPN que *personne* Gaël Monfils n'était pas assez professionnel. « *personne* Monfils aurait déjà dû gagner 4 ou 5 majeurs », a-t-il précisé.

Entités nommées

Normalisation

personne *date*

Le joueur de tennis américain John McEnroe a déclaré samedi

organisation *personne*

sur ESPN que Gaël Monfils n'était pas assez professionnel.

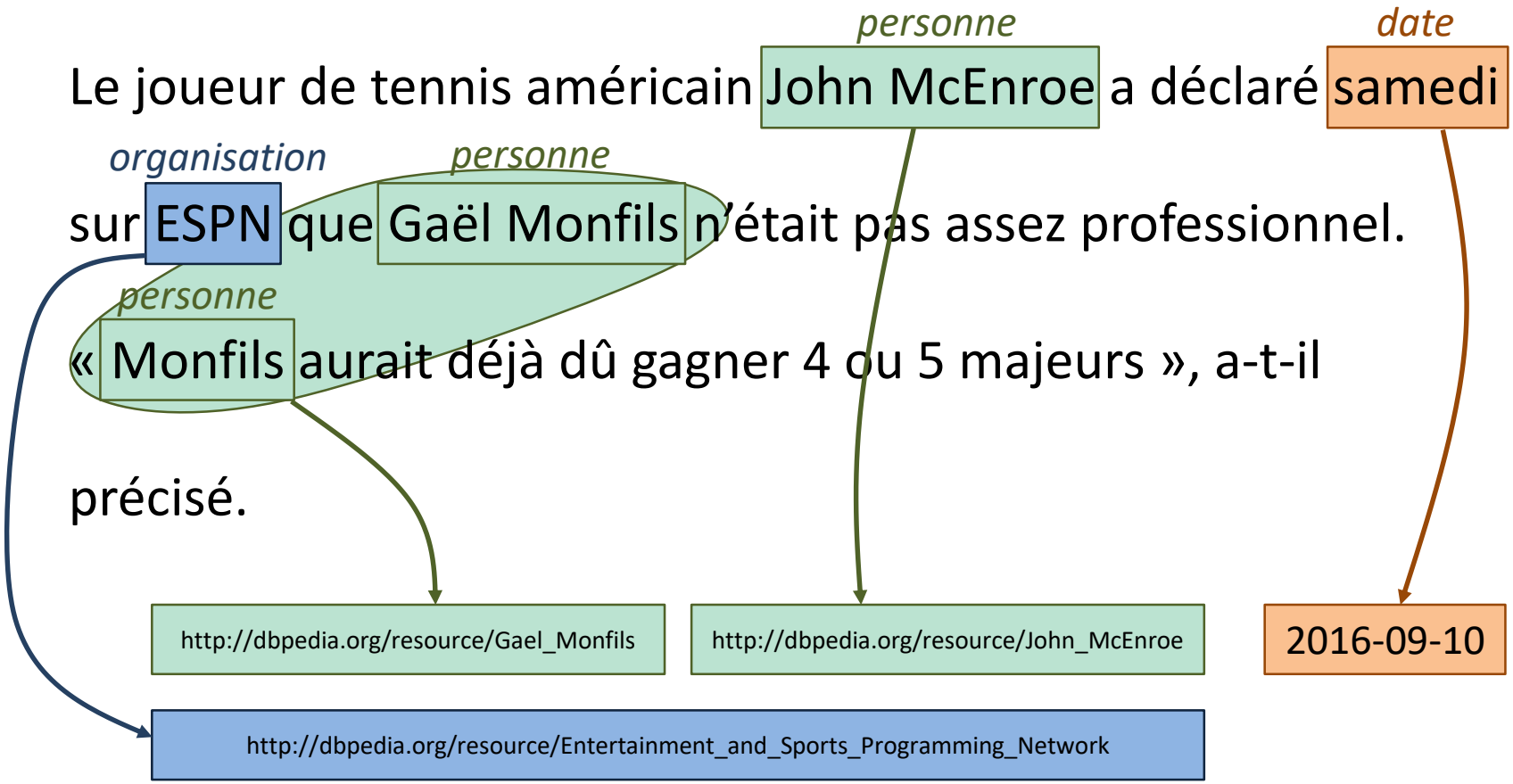
personne

« Monfils aurait déjà dû gagner 4 ou 5 majeurs », a-t-il

précisé.

Entités nommées

Normalisation



Entités nommées

Plus de précision ?

fonction *pers:sportif* *date:jour*

Le joueur de tennis américain John McEnroe a déclaré samedi

org:TV *pers:sportif*

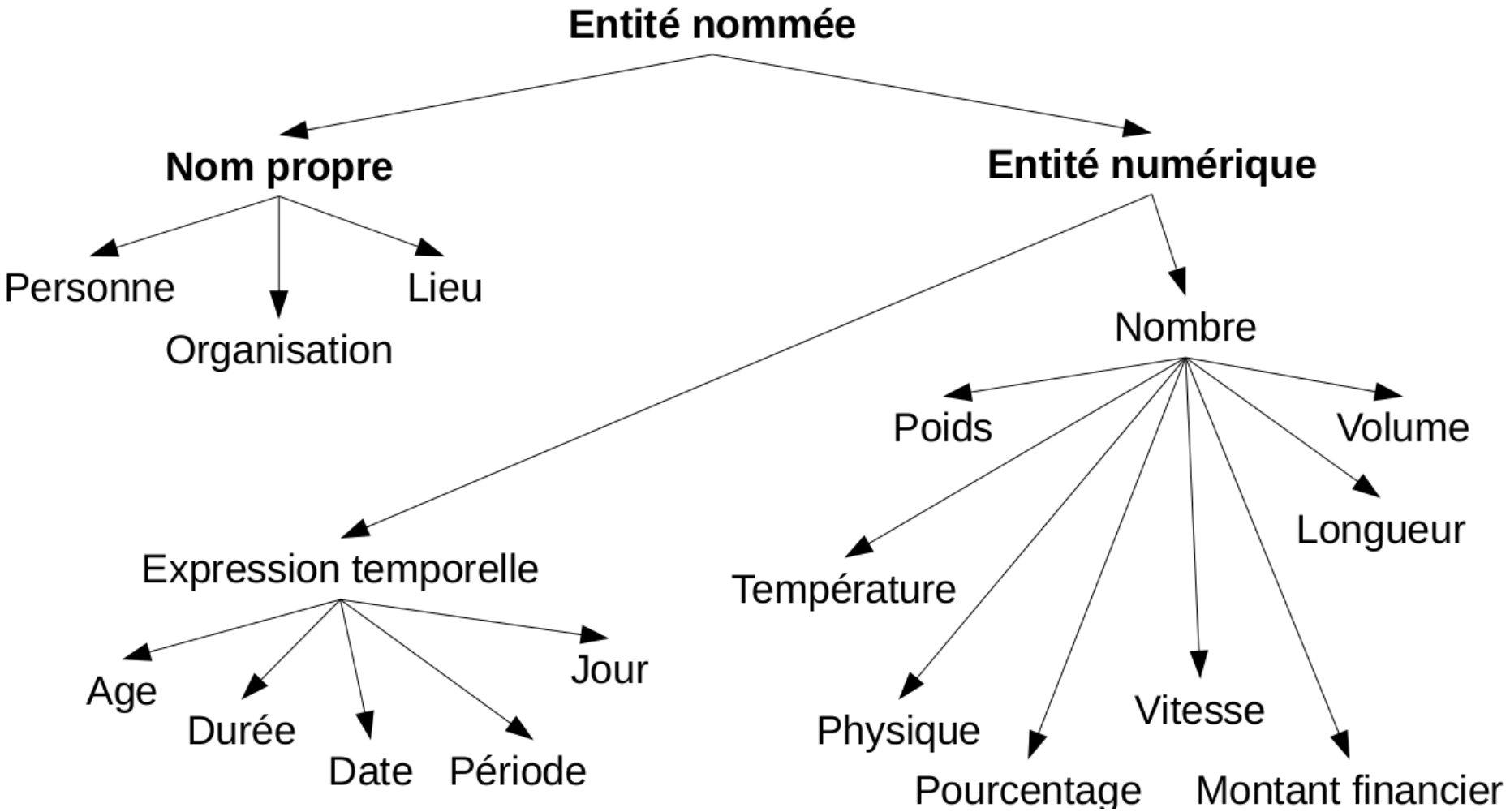
sur ESPN que Gaël Monfils n'était pas assez professionnel.

pers:sportif *evt:tournoi*

« Monfils aurait déjà dû gagner Rolland-Garros », a-t-il

précisé.

Entités nommées : un exemple de hiérarchie



Intérêt des entités nommées

- Les entités nommées peuvent être indexées, liées entre elles, etc.
- Des caractéristiques ou des relations peuvent être extraites
- Les entités nommées sont souvent la réponse à des besoins d'information (questions factuelles)

- Usage interne : les entités nommées peuvent aider d'autres tâches de traitement automatique des langues (traduction, analyse syntaxique)

- Particularité des entités nommées : ce sont des séquences de mots

Modèles en séquences

Texte	*Format BIO*
Le	O
joueur	O
de	O
tennis	O
américain	O
John	B-PERS
McEnroe	I-PERS
a	O
déclaré	O
samedi	B-DATE
sur	O
ESPN	B-ORG
que	O
Gaël	B-PERS
Monfils	I-PERS

Modèles en séquences

Texte	*Format BIO*	*Format IO*
Le	O	O
joueur	O	O
de	O	O
tennis	O	O
américain	O	O
John	B-PERS	I-PERS
McEnroe	I-PERS	I-PERS
a	O	O
déclaré	O	O
samedi	B-DATE	I-DATE
sur	O	O
ESPN	B-ORG	I-ORG
que	O	O
Gaël	B-PERS	I-PERS
Monfils	I-PERS	I-PERS

Format BIO meilleure représentation mais plus gourmand (plus lent).
Pas toujours de meilleures performances en pratique

Autres conventions : BIOE, IOBES…
Ce sont des classes au sens classique du terme.

Modèles en séquences

Modèles renommés pour la prédiction de séquences :

- MEMM : Maximum Entropy Markov Models
- CRF : Conditional Random Fields
- Réseaux de neurones récurrents (ex : LSTM, bi-LSTM, etc.)
- Représentation contextuelle (ex: BERT)
- Exhaustive biaffine

Modèles à base de « feature engineering »

Modèles en séquences : les traits

Texte	*Format IO*
Le	O
joueur	O
de	O
tennis	O
américain	O
John	I-PERS
McEnroe	I-PERS
a	O
déclaré	O
samedi	I-DATE
sur	O
ESPN	I-ORG
que	O
Gaël	I-PERS
Monfils	I-PERS

Les traits (*features*) concernent une ligne, soit un élément de la séquence.

Modèles en séquences : les traits

Texte	*Format IO*
Le	O
joueur	O
de	O
tennis	O
américain	O
John	I-PERS
McEnroe	I-PERS
a	O
déclaré	O
samedi	I-DATE
sur	O
ESPN	I-ORG
que	O
Gaël	I-PERS
Monfils	I-PERS

Les traits (*features*) concernent une ligne, soit un élément de la séquence.

le mot

Modèles en séquences : les traits

Texte	*Format IO*
Le	O
joueur	O
de	O
tennis	O
américain	O
John	I-PERS
McEnroe	I-PERS
a	O
déclaré	O
samedi	I-DATE
sur	O
ESPN	I-ORG
que	O
Gaël	I-PERS
Monfils	I-PERS

Les traits (*features*) concernent une ligne, soit un élément de la séquence.

le mot

les mots suivantes/précédents

Modèles en séquences : les traits

Texte	*Format IO*
Le	O
joueur	O
de	O
tennis	O
américain	O
John	I-PERS
McEnroe	I-PERS
a	O
déclaré	O
samedi	I-DATE
sur	O
ESPN	I-ORG
que	O
Gaël	I-PERS
Monfils	I-PERS

Les traits (*features*) concernent une ligne, soit un élément de la séquence.

- le mot
- les mots suivantes/précédents
- la classe précédente (et parfois suivante)

Modèles en séquences : les traits

Texte	*Format IO*
Le	O
joueur	O
de	O
tennis	O
américain	O
John	I-PERS
McEnroe	I-PERS
a	O
déclaré	O
samedi	I-DATE
sur	O
ESPN	I-ORG
que	O
Gaël	I-PERS
Monfils	I-PERS

Les traits (*features*) concernent une ligne, soit un élément de la séquence.

- le mot
- les mots suivantes/précédents
- la classe précédente (et parfois suivante)
- d'autres traits

Modèles en séquences : les traits

Texte	*POS*	*Format IO*
Le	DET	O
joueur	NOM	O
de	PREP	O
tennis	NOM	O
américain	ADJ	O
John	NP	I-PERS
McEnroe	NP	I-PERS
a	V-AUX	O
déclaré	V-PPAS	O
samedi	NOM	I-DATE
sur	PREP	O
ESPN	NP	I-ORG
que	CONJ	O
Gaël	NP	I-PERS
Monfils	NP	I-PERS

Étiquettes morpho-syntaxiques (POS)

Modèles en séquences : les traits

Texte	*POS*	*MAJ*	*Format IO*
Le	DET	YES	O
joueur	NOM	NO	O
de	PREP	NO	
tennis	NOM	NO	O
américain	ADJ	NO	O
John	NP	YES	I-PERS
McEnroe	NP	YES	I-PERS
a	V-AUX	NO	O
déclaré	V-PPAS	NO	O
samedi	NOM	NO	I-DATE
sur	PREP	NO	O
ESPN	NP	YES	I-ORG
que	CONJ	NO	O
Gaël	NP	YES	I-PERS
Monfils	NP	YES	I-PERS

Commence par une majuscule ?

Mais aussi :
- contient une sous-chaîne précise
- préfixe/suffixe
- taille
- contient des chiffres/alphabet grec/ponctuation
- ne contient que des chiffres/alphabet grec/ponctuation
- …

Modèles en séquences : les traits

Texte	*POS*	*MAJ*	*SPEECH_VERB*	*Format IO*
Le	DET	YES	NO	O
joueur	NOM	NO	NO	O
de	PREP	NO	NO	O
tenn			NO	O
amé			NO	O
John	NP	YES	NO	I-PERS
McEnroe	NP	YES	NO	I-PERS
a	V-AUX	NO	NO	O
déclaré	V-PPAS	NO	**YES**	O
samedi	NOM	NO	NO	I-DATE
sur	PREP	NO	NO	O
ESPN	NP	YES	NO	I-ORG
que	CONJ	NO	NO	O
Gaël	NP	YES	NO	I-PERS
Monfils	NP	YES	NO	I-PERS

Fait partie d'un lexique ?

Modèles en séquences : les traits

Texte	*POS*	*MAJ*	*SPEECH_VERB*	*FIRST_NAME*	*Format IO*
Le	DET	YES	NO	NO	O
joueur	NOM	NO	NO	NO	O
de	PREP	NO	NO	NO	O
tenn			NO	NO	O
amé			NO	NO	O
John	NP	YES	NO	**YES**	I-PERS
McEnroe	NP	YES	NO	NO	I-PERS
a	V-AUX	NO	NO	NO	O
déclaré	V-PPAS	NO	YES	NO	O
samedi	NOM	NO	NO	NO	I-DATE
sur	PREP	NO	NO	NO	O
ESPN	NP	YES	NO	NO	I-ORG
que	CONJ	NO	NO	NO	O
Gaël	NP	YES	NO	NO	I-PERS
Monfils	NP	YES	NO	NO	I-PERS

Fait partie d'un lexique ?

Modèles en séquences : les traits

Texte	*POS*	*MAJ*	*SPEECH_VERB*	*FIRST_NAME*	*Format IO*
Le	DET	YES	NO	NO	O
joueur	NOM			NO	O
de	PREP			NO	O
tennis	NOM	NO	NO	NO	O
américain	ADJ	NO	NO	NO	O
John	NP	YES	NO	YES	I-PERS
McEnroe	NP	YES	NO	NO	I-PERS
a	V-AUX	NO	NO	NO	O
déclaré	V-PPAS	NO	YES	NO	O
samedi	NOM	NO	NO	NO	I-DATE
sur	PREP	NO	NO	NO	O
ESPN	NP	YES	NO	NO	I-ORG
que	CONJ	NO	NO	NO	O
Gaël	NP	YES	NO	NO	I-PERS
Monfils	NP	YES	NO	NO	I-PERS

Bigrammes, n-grammes

Modèles en séquences : les traits

Texte	*POS*	*MAJ*	*SPEECH_VERB*	*FIRST_NAME*	*Format IO*
Le	DET	YES	NO	NO	O
joueur	NOM			NO	O
de	PREP			NO	O
tennis	NOM	NO	NO	NO	O
américain	ADJ	NO	NO	NO	O
John	NP	YES	NO	YES	I-PERS
McEnroe	NP	YES	NO	NO	I-PERS
a	V-AUX	NO	NO	NO	O
déclaré	V-PPAS	NO	YES	NO	O
samedi	NOM	NO	NO	NO	I-DATE
sur	PREP	NO	NO	NO	O
ESPN	NP	YES	NO	NO	I-ORG
que	CONJ	NO	NO	NO	O
Gaël	NP	YES	NO	NO	I-PERS
Monfils	NP	YES	NO	NO	I-PERS

Bigrammes, n-grammes

Modèles en séquences : les traits

Texte	*POS*	*MAJ*	*SPEECH_VERB*	*FIRST_NAME*	*Format IO*
Le	DET	YES	NO	NO	O
joueur	NOM	NO	NO	NO	O
de	PREP	NO	NO	NO	O
tennis	NOM	NO	NO	NO	O
américain	ADJ	NO	NO	NO	O
John	NP	YES	NO	YES	I-PERS
McEnroe	NP	YES	NO	NO	I-PERS
a	V-AUX	NO	NO	NO	O
déclaré	V-PPAS	NO	YES	NO	O
samedi	NOM	NO	NO	NO	I-DATE
sur	PREP	NO	NO	NO	O
ESPN	NP	YES	NO	NO	I-ORG
que	CONJ	NO	NO	NO	O
Gaël	NP	YES	NO	NO	I-PERS
Monfils	NP	YES	NO	NO	I-PERS

Bigrammes, n-grammes

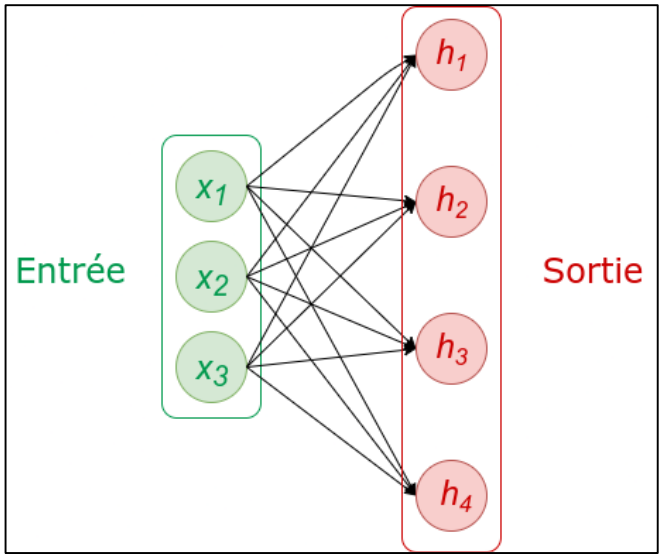
SORBONNE UNIVERSITÉ

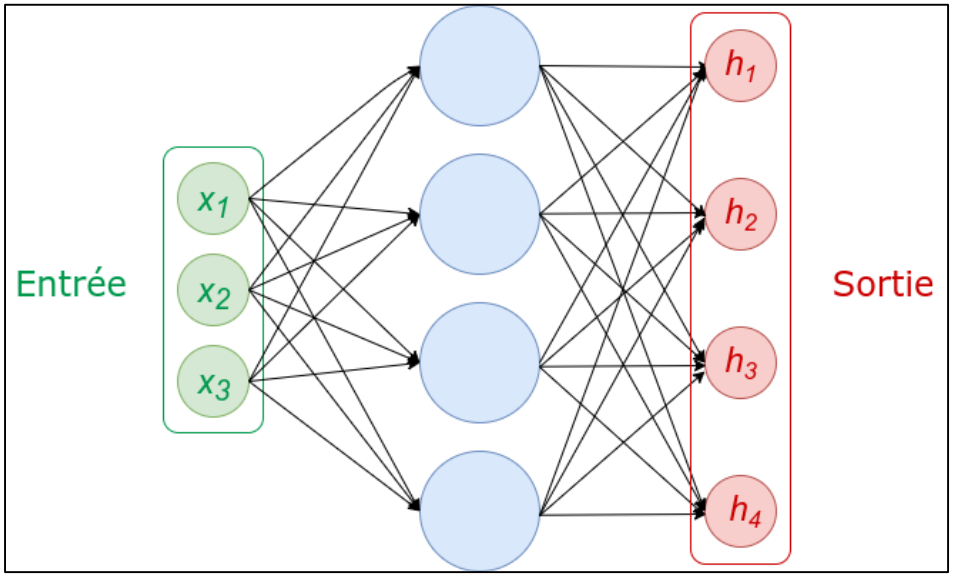
Introduction aux réseaux de neurones récurrents

Master DAC, Sorbonne Université

Xavier Tannier
xavier.tannier@sorbonne-universite.fr

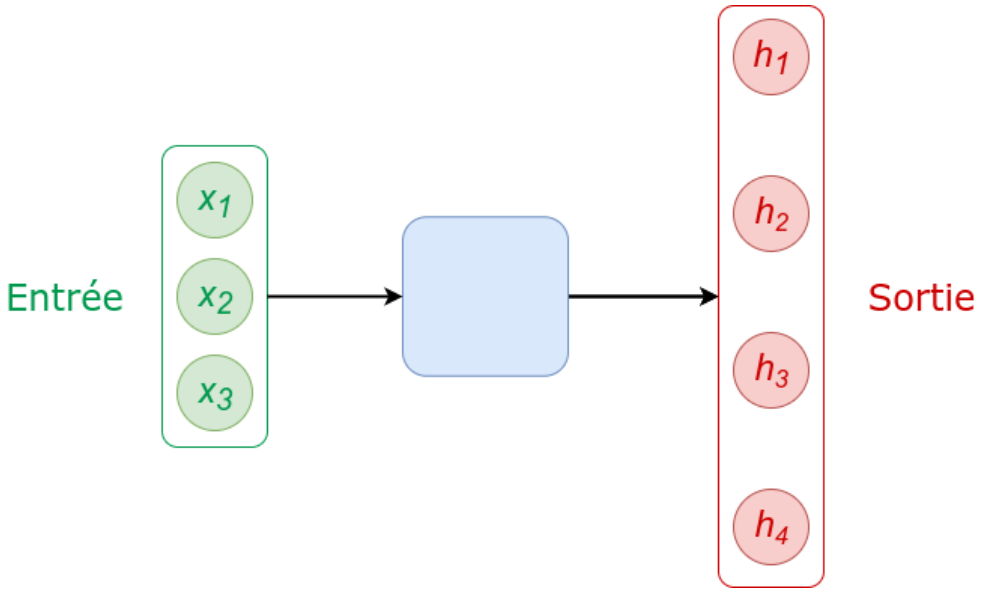
Un réseau statique





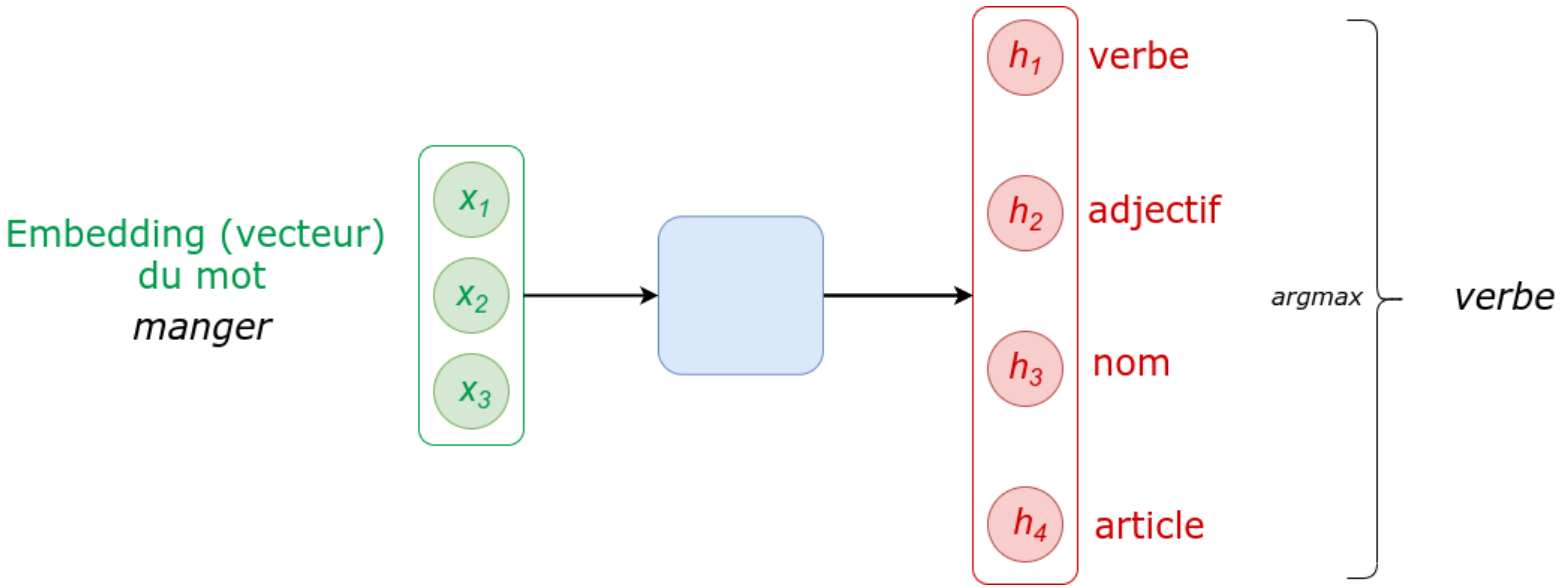
etc...

Un réseau statique



Un réseau statique

- Exemple :
  prédire si un mot est un verbe, un adjectif, un nom ou une préposition



Un réseau statique

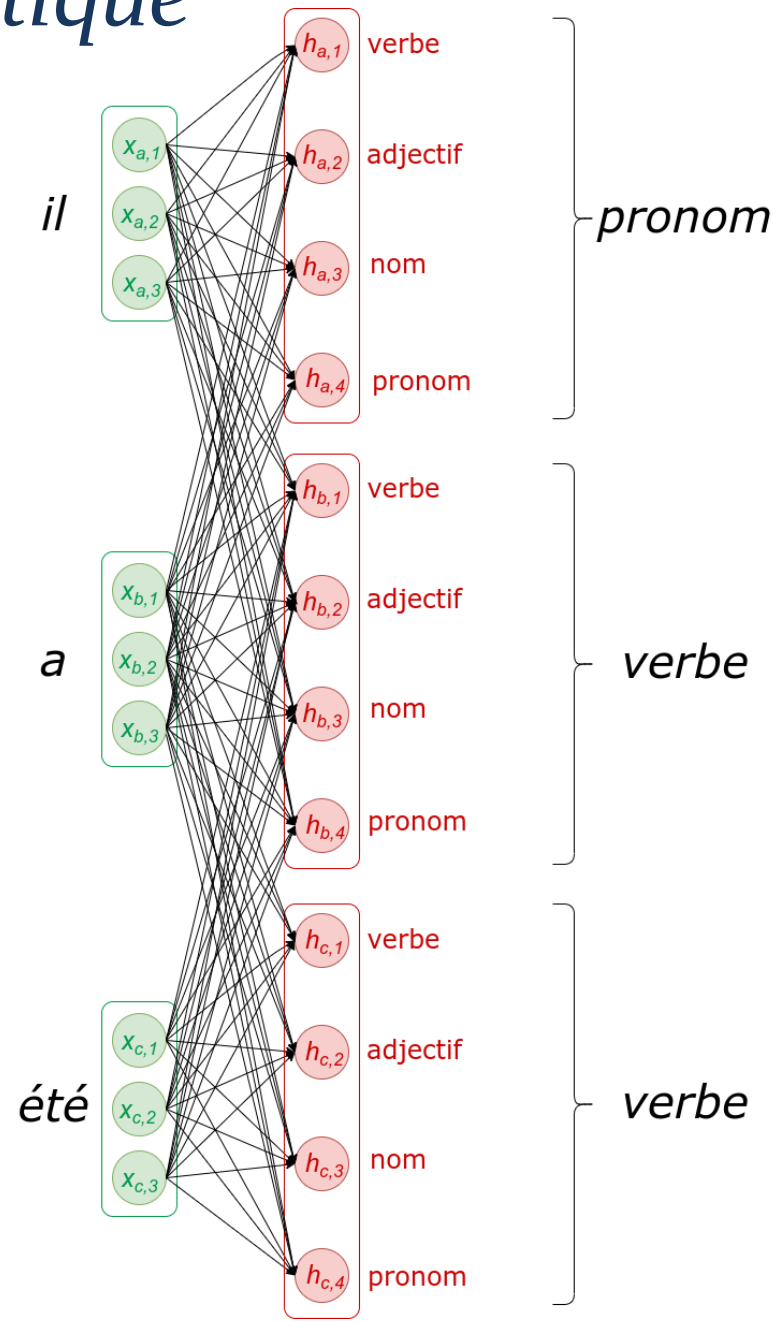
Problèmes :

1. La catégorie d'un mot dépend du contexte de ce mot.
Ce n'est pas une prédiction indépendante de la prédiction des mots autour
(« il a *été*$_{Verbe}$ élu à la direction » vs « il est parti tout l'*été*$_{Nom}$ »)

2. On veut parfois prédire des classes pour des groupes de mots (*n-grammes*)
(« *[Marie Curie]*$_{Personne}$ a obtenu le *[Prix Nobel de Chimie]*$_{Distinction}$ »)

Un réseau statique

Concaténation des vecteurs de mots ?



Un réseau statique

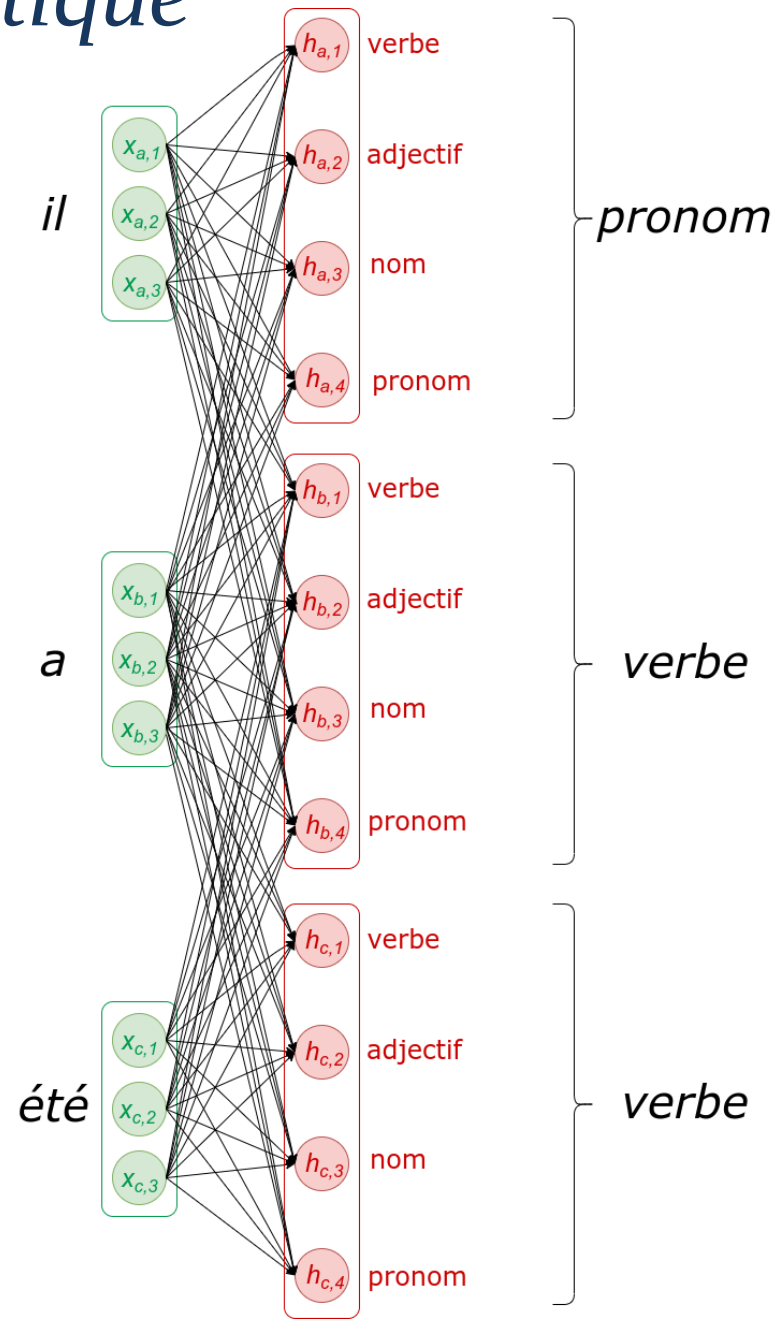
Concaténation des vecteurs de mots ?

Mais :

3. On obtient un très gros réseau
   (avec beaucoup de paramètres)

4. Les phrases ont des tailles variables
   (et un réseau doit avoir
   une taille fixe et prédéfinie)

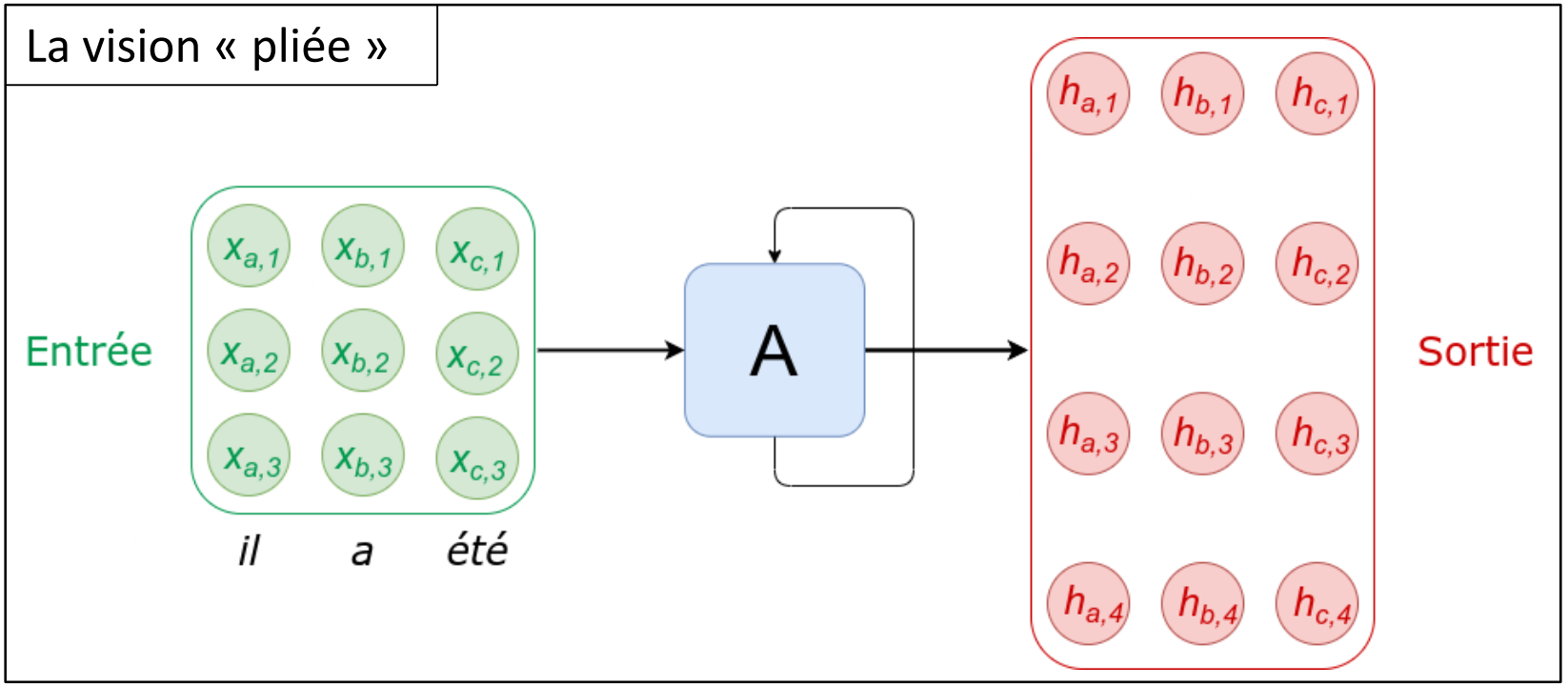
5. Les mots peuvent être à des positions différentes dans la phrase
   (et les poids d'un réseau
   ne se déplacent pas, eux)



Réseaux de Neurones Récurrents (RNN)

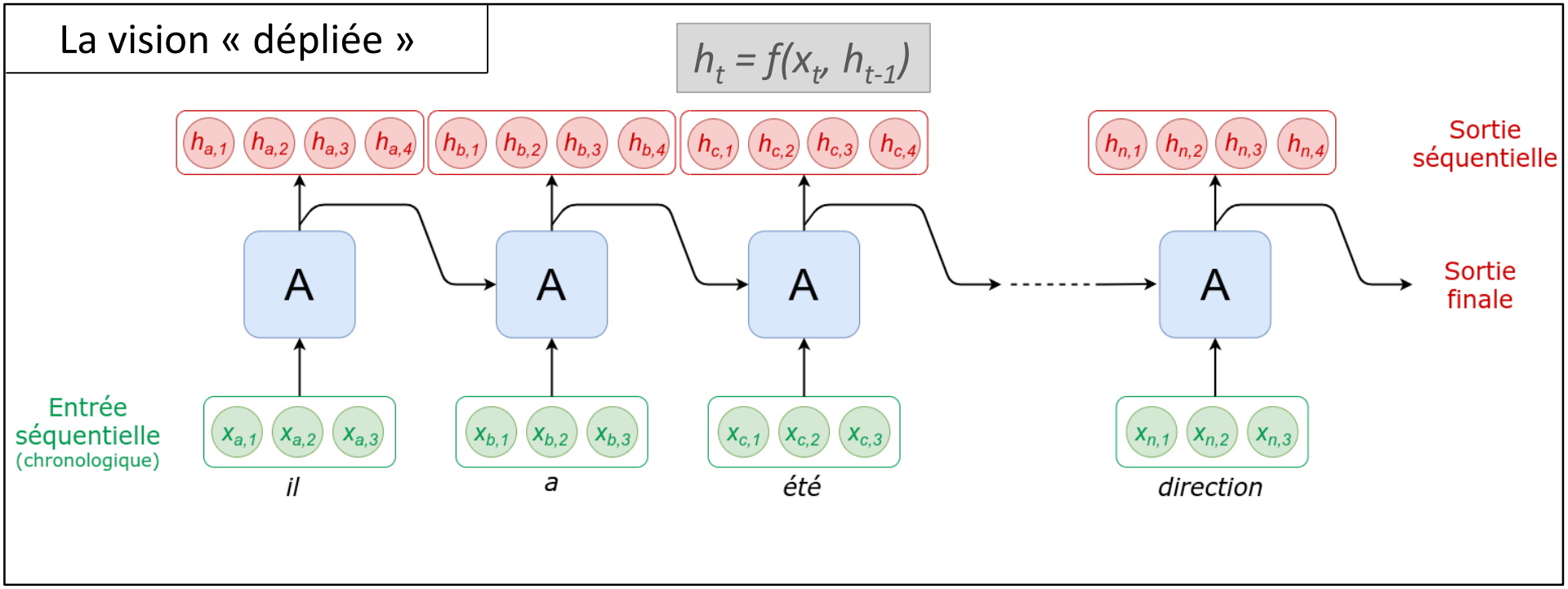
Réseau récurrent

Une solution possible : le réseau récurrent

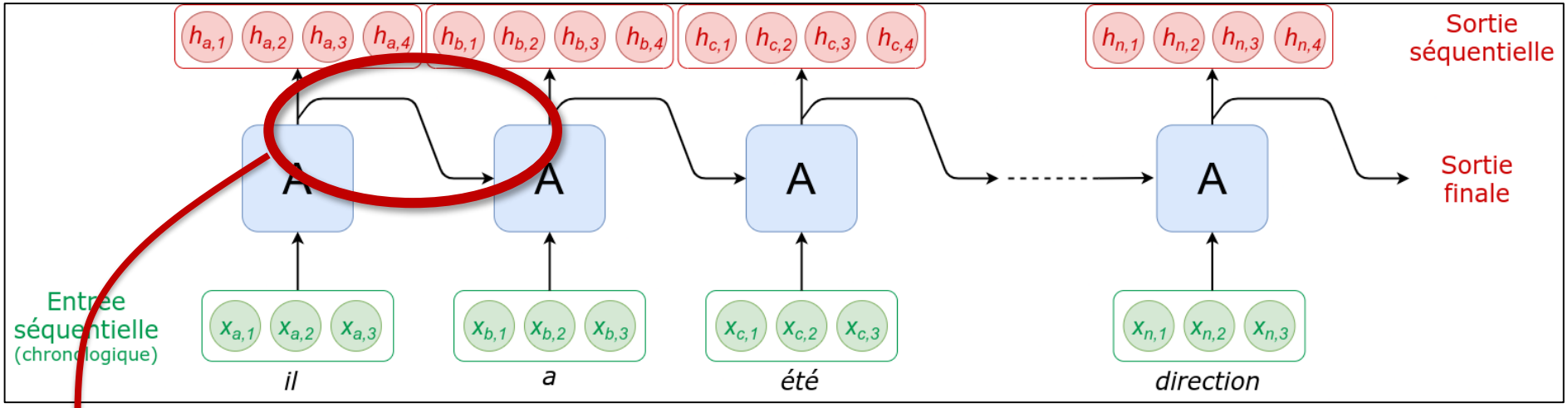


Réseau récurrent

Une solution possible : le réseau récurrent

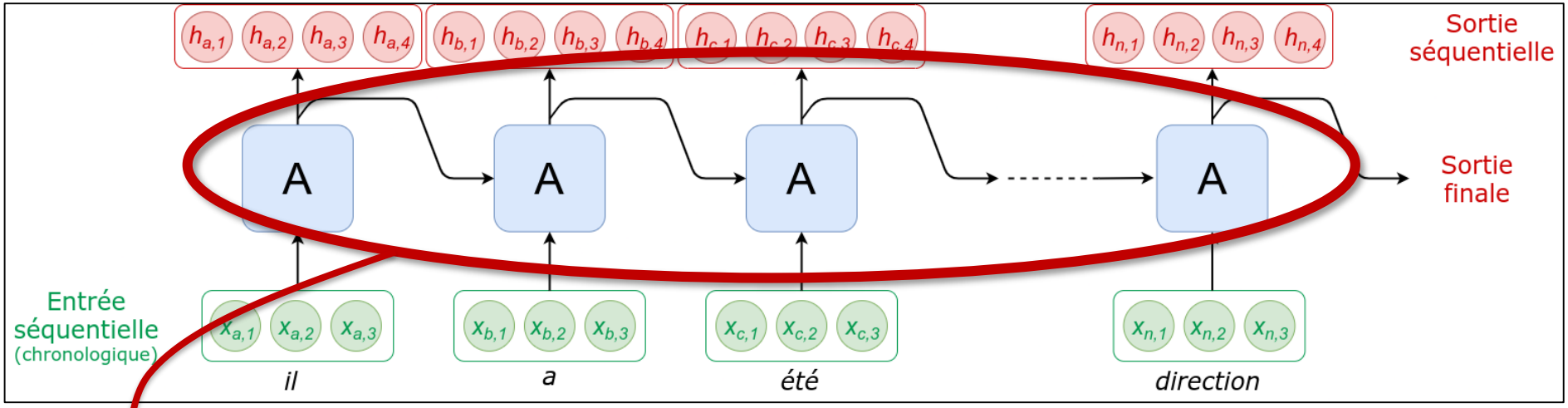


Réseau récurrent



Récurrence → 1. La catégorie d'un mot dépend du contexte de ce mot

Réseau récurrent



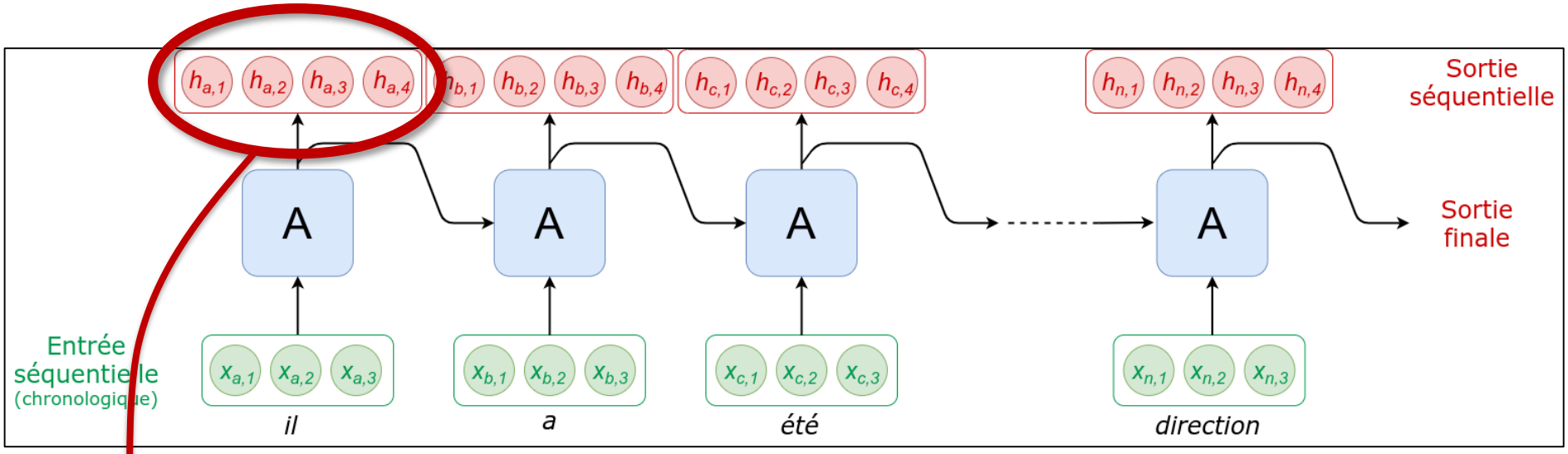
Une seule unité récurrente

3. Taille du réseau

4. Longueur variable des phrases

5. Position variable des mots

Réseau récurrent

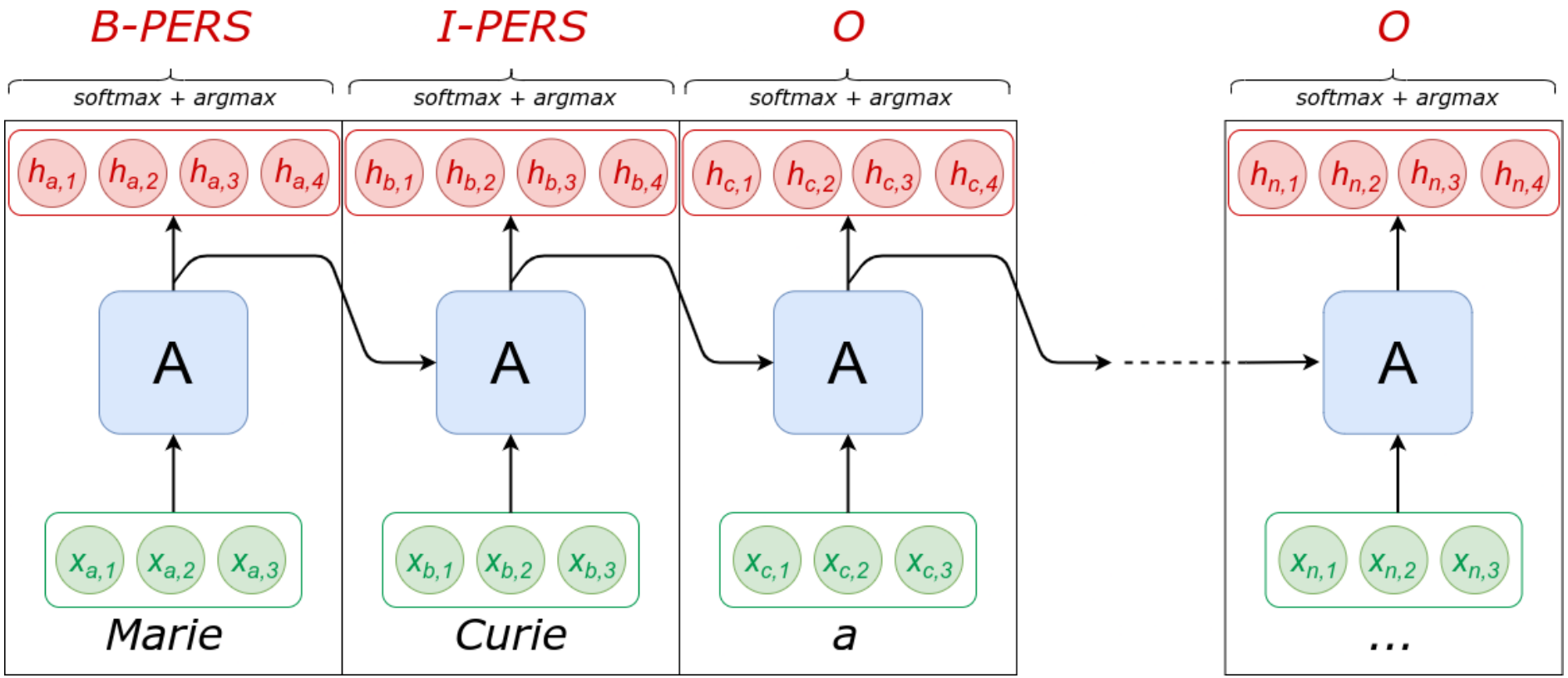


Une sortie par token

2. On peut prédire des classes
pour des groupes de mots
(grâce au format IO, BIO, IOBES, etc.)

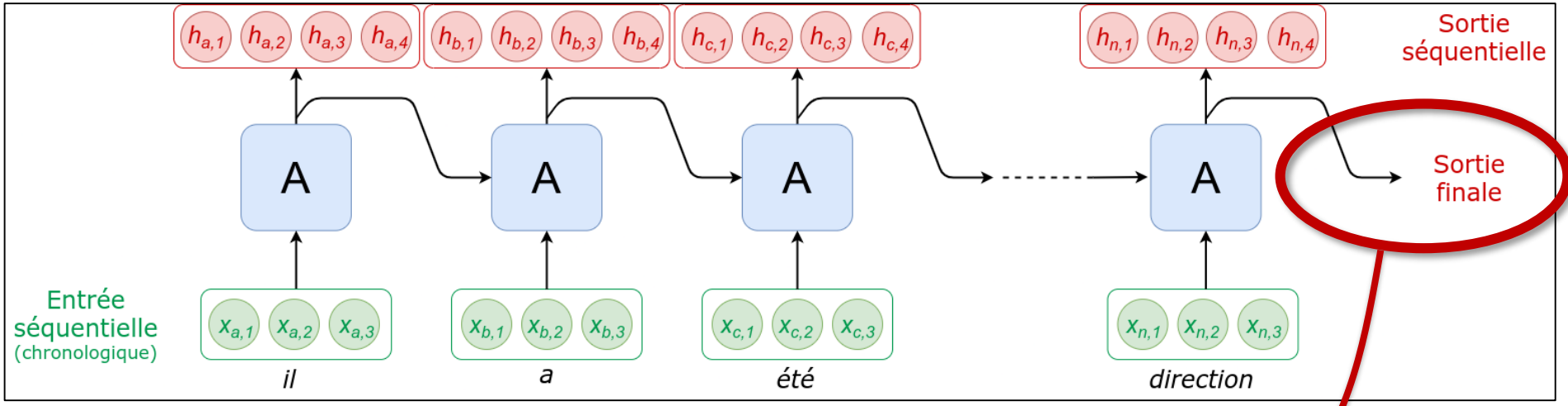
Réseau récurrent

Pour de la reconnaissance d'entités nommées



Réseau récurrent

Pour de la classification de texte

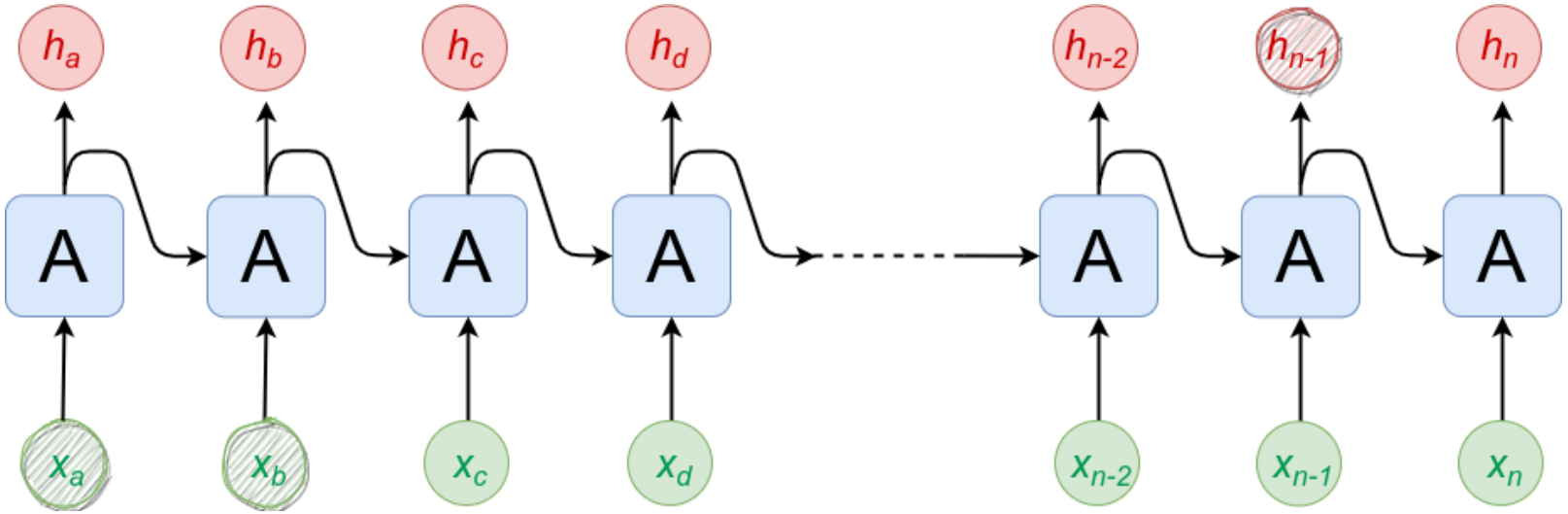


Utilisation de la sortir finale pour faire une classification de la séquence complète

Long Short-Term Memory Networks (LSTM)

LSTM

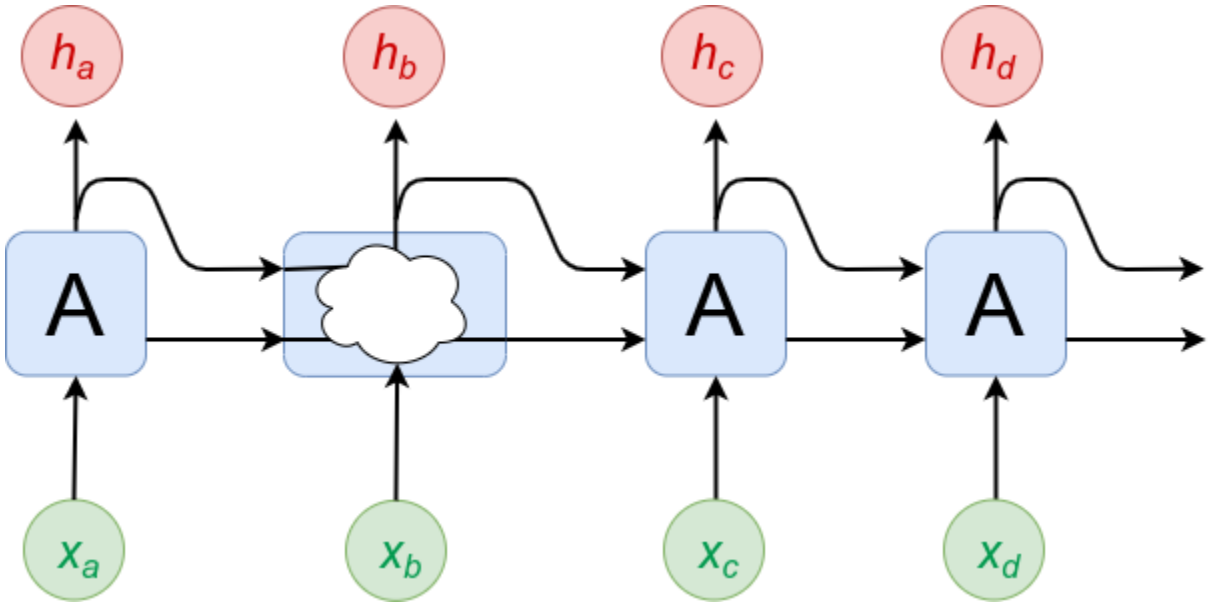
- Les réseaux récurrents basiques ont du mal à « retenir » les informations contextuelles utiles



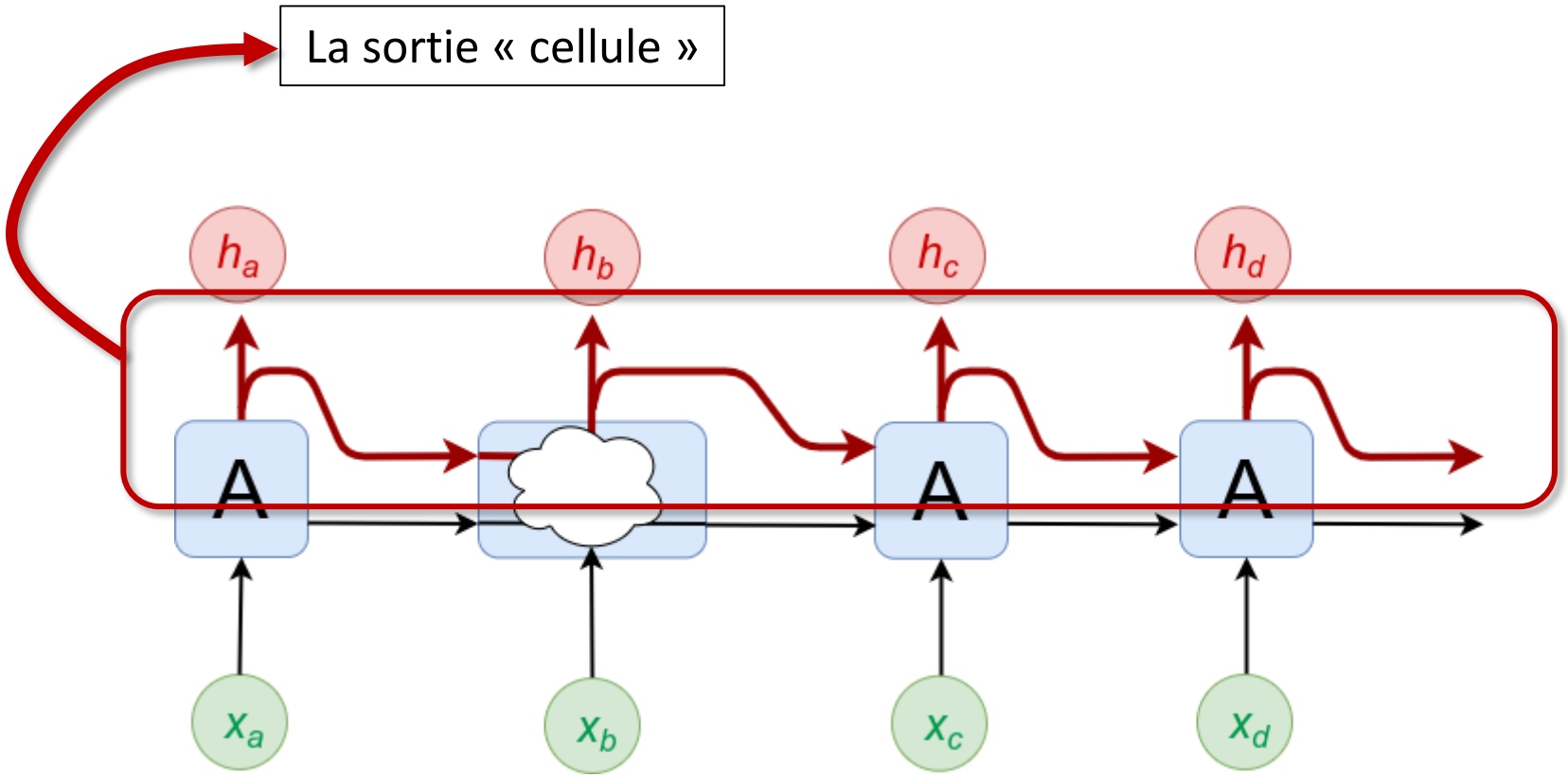
- En théorie, les RNN en sont capables (il existe un chemin de x_a à h_{n-1})
- En pratique c'est difficile

LSTM

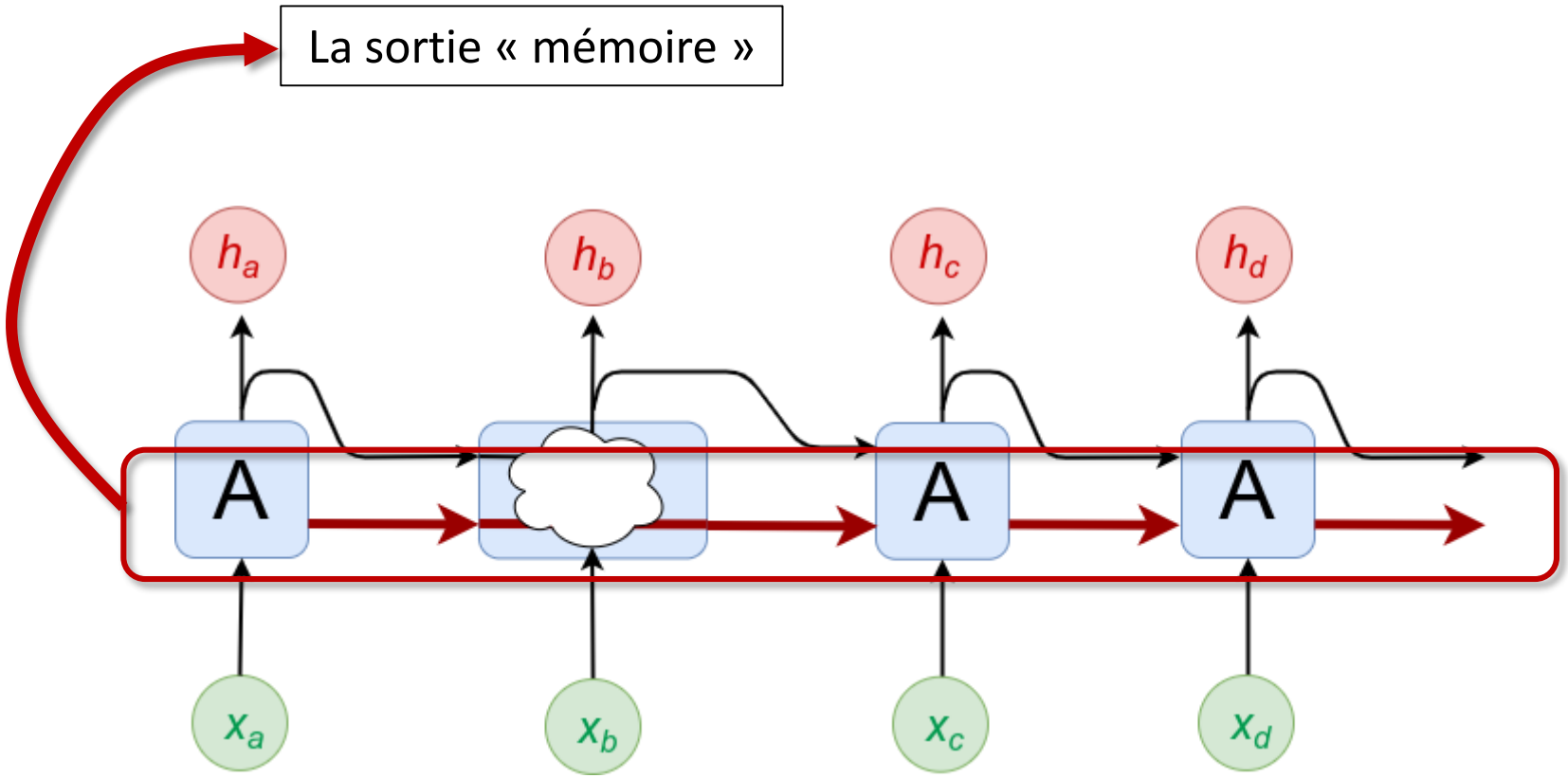
- Idée : ajouter une nouvelle entrée/sortie à chaque unité, dont le but sera de se « souvenir » des choses importantes



LSTM

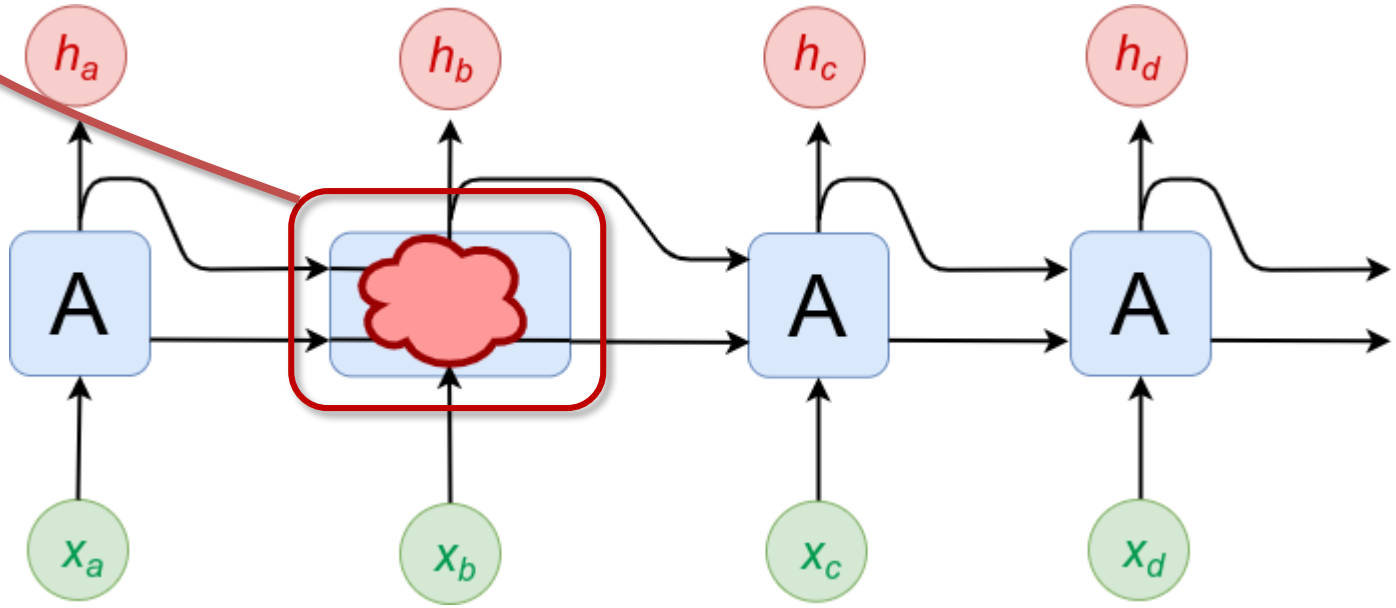


LSTM



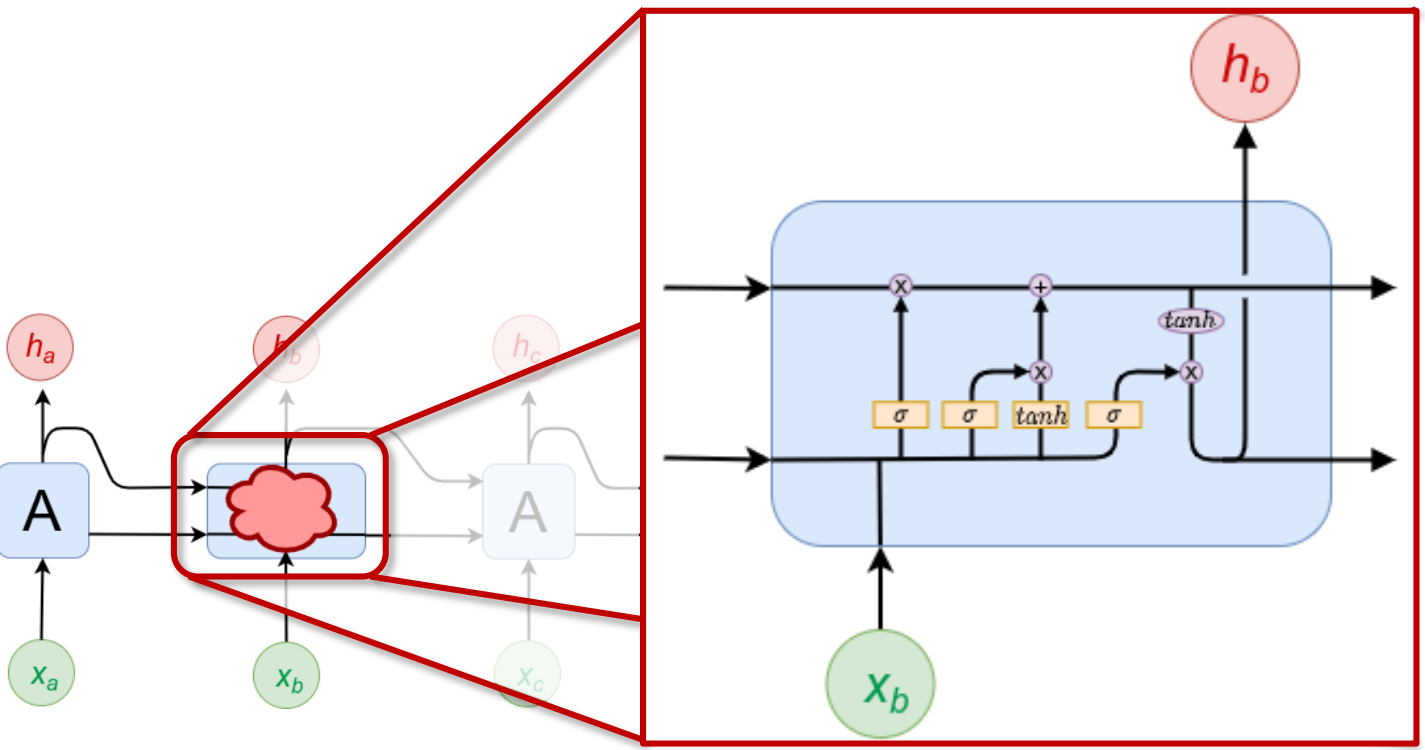
LSTM

Interactions entre les trois entrées et les deux sorties



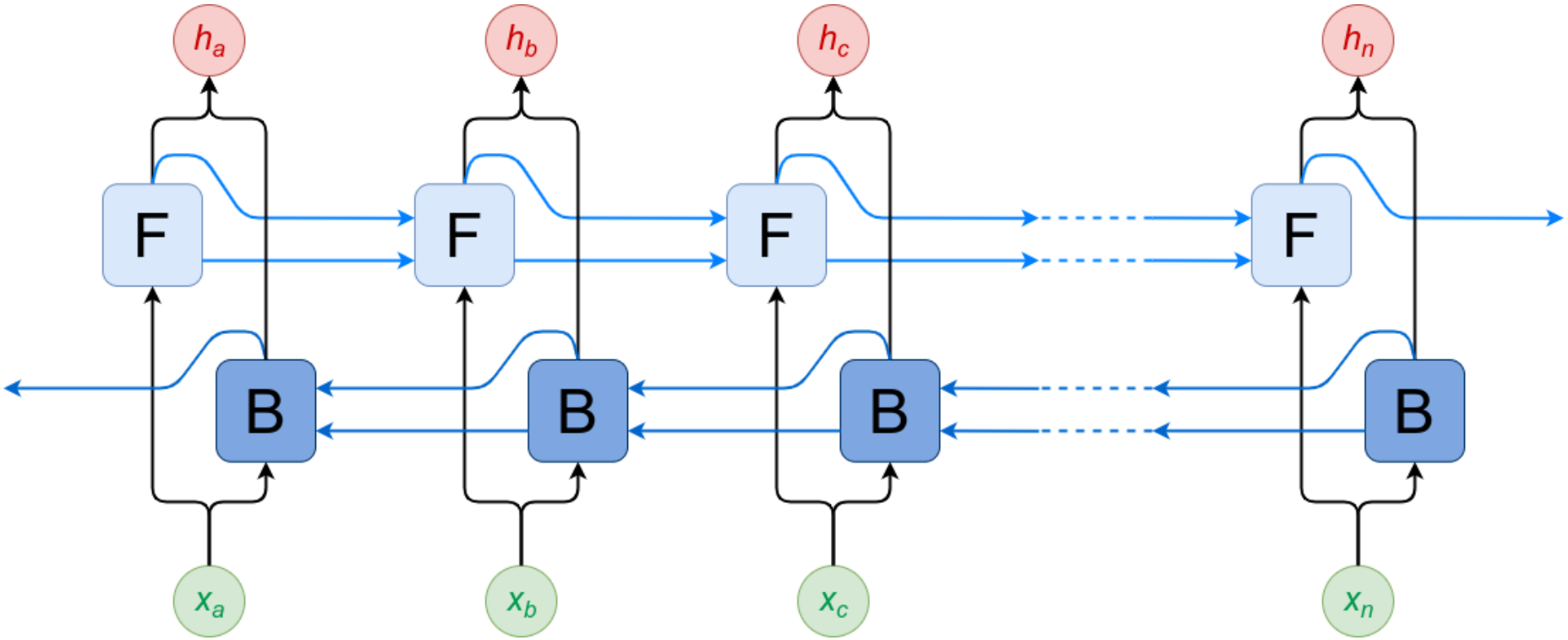
LSTM

Interactions entre les trois entrées et les deux sorties :
- oubli explicite des informations non pertinentes (forget gate)
- ajout d'information sur le nouveau token (update gate)
- construction de la sortie (output gate)



Bi-LSTM

- Parcourir le texte dans les deux sens
- Une cellule profite du contexte à gauche et à droite



LSTM

Remarques :

- Le réseau récurrent n'est qu'un modèle de représentation de l'entrée, pas un modèle de classification.
- On peut brancher n'importe quel modèle de classification sur un LSTM
- On peut brancher n'importe quel modèle de représentation de séquences sur un classifieur
- On peut superposer / concaténer plusieurs modèles de représentation (modèles de caractères, embeddings contextuels, features précalculées, etc.)

LSTM

- Plus de détails sur l'architecture LSTM :
  https://colah.github.io/posts/2015-08-Understanding-LSTMs/

- Beaucoup plus détails sur l'architecture LSTM :

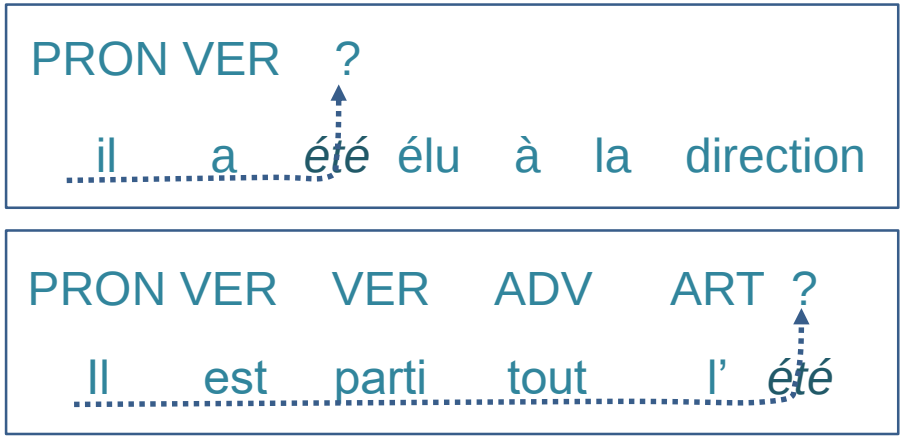
  Understanding LSTM – a tutorial into Long Short-Term Memory Recurrent Neural Networks
  Ralf C. Staudemeyer, Eric Rothstein Morris
  https://arxiv.org/pdf/1909.09586.pdf

- Voir aussi les implémentations en tensorflow, pytorch, etc.

Champs Aléatoires Conditionnels (CRF)

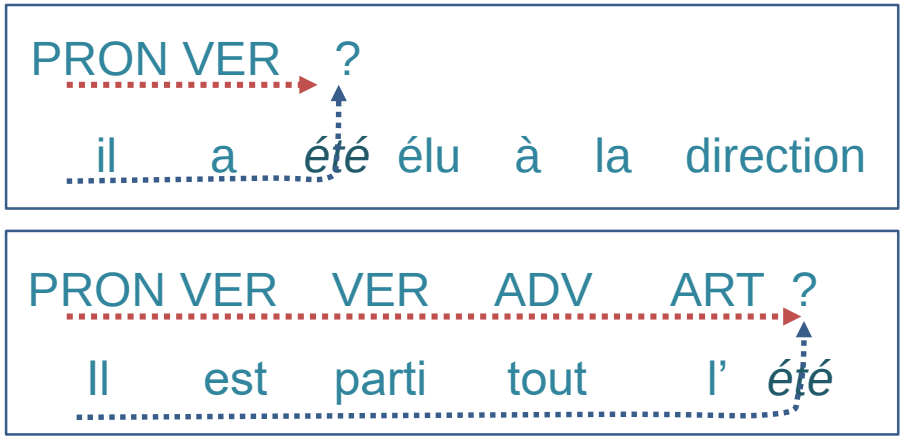
CRF

- Les modèles vus précédemment calculent les scores des classes uniquement en fonction de la représentation des mots (« scores d'émission »)



- Pourtant, la particularité d'un problème de séquences est que les classes sont interdépendantes.

  Le CRF permet de calculer les « scores de transition »



CRF

- Score de transition : vraisemblance qu'un mot ait telle classe sachant la classe du mot précédent
- Un tenseur de taille |V| x n x n

Token : « été »

Classe à prédire

Classe précédente	<start>	pronom	verbe	nom	<end>
<start>	-10000	-2.23	1.54	2.98	-10000
pronom	-10000	-4.85	2.07	-0.46	-2.48
verbe	-10000	-5.59	3.89	-2.90	-1.92
nom	-10000	-3.52	-0.78	0.36	-2.45
<end>	-10000	-10000	-10000	-10000	-10000

V = vocabulaire
n = nombre de classes

CRF

- La prédiction est la séquence qui maximise la vraisemblance globale
  (une version discriminante des champs aléatoires de Markov)

- Un CRF est un modèle de classification branché sur une représentation (pas forcément neuronale)

- Dans le format BIO, le CRF permet notamment d'empêcher facilement la prédiction de séquence illicites (ex : **`B-PERS I-LOC`**)

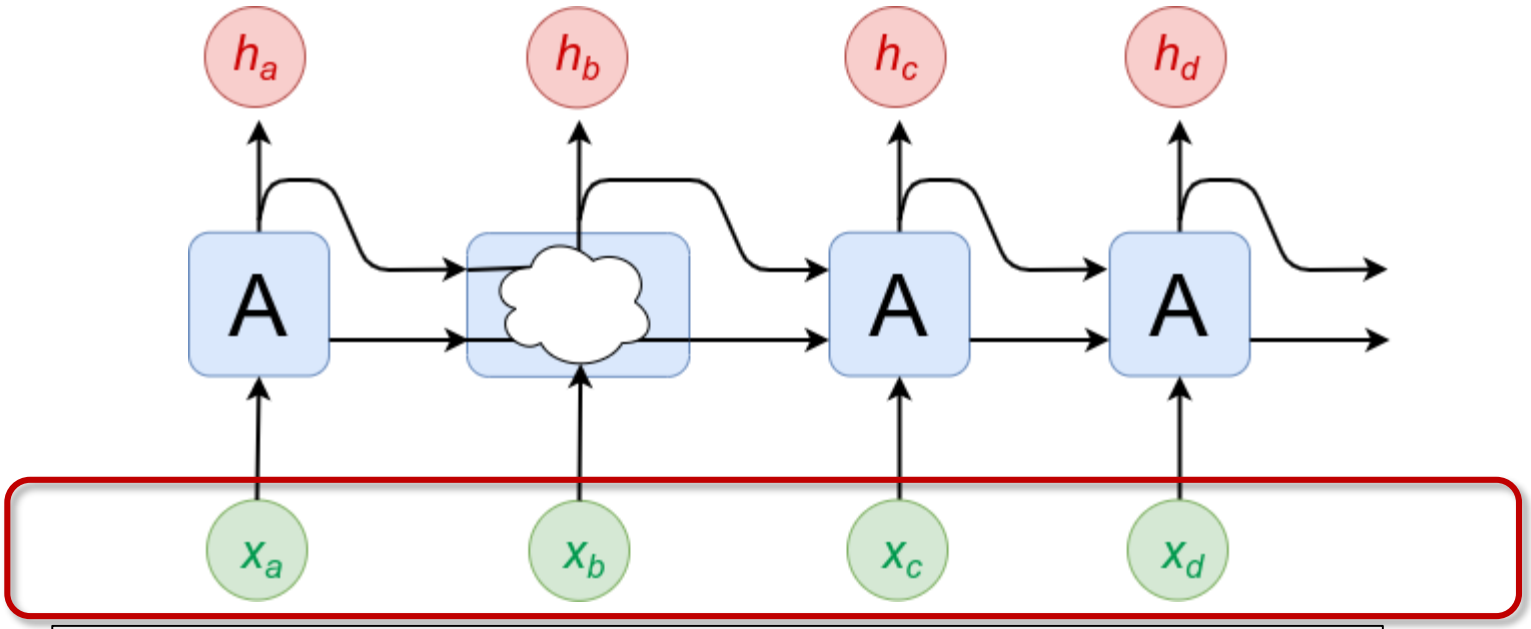
CRF

- Plus de détails :
  - https://towardsdatascience.com/conditional-random-fields-explained-e5b8256da776
  - Différences entre HMM, MEMM et CRF : https://www.alibabacloud.com/blog/hmm-memm-and-crf-a-comparative-analysis-of-statistical-modeling-methods_592049

- Beaucoup plus de détails :
  - An introduction to Conditional Random Fields https://www.pure.ed.ac.uk/ws/portalfiles/portal/10482724/crftut_fnt.pdf
  - Conditional random fields: Probabilistic models for segmenting and labeling sequence data, Lafferty, J., McCallum, A., Pereira, F., Proc. 18th International Conf. on Machine Learning, Morgan Kaufmann, p. 282–289, 2001

Un peu de pratique

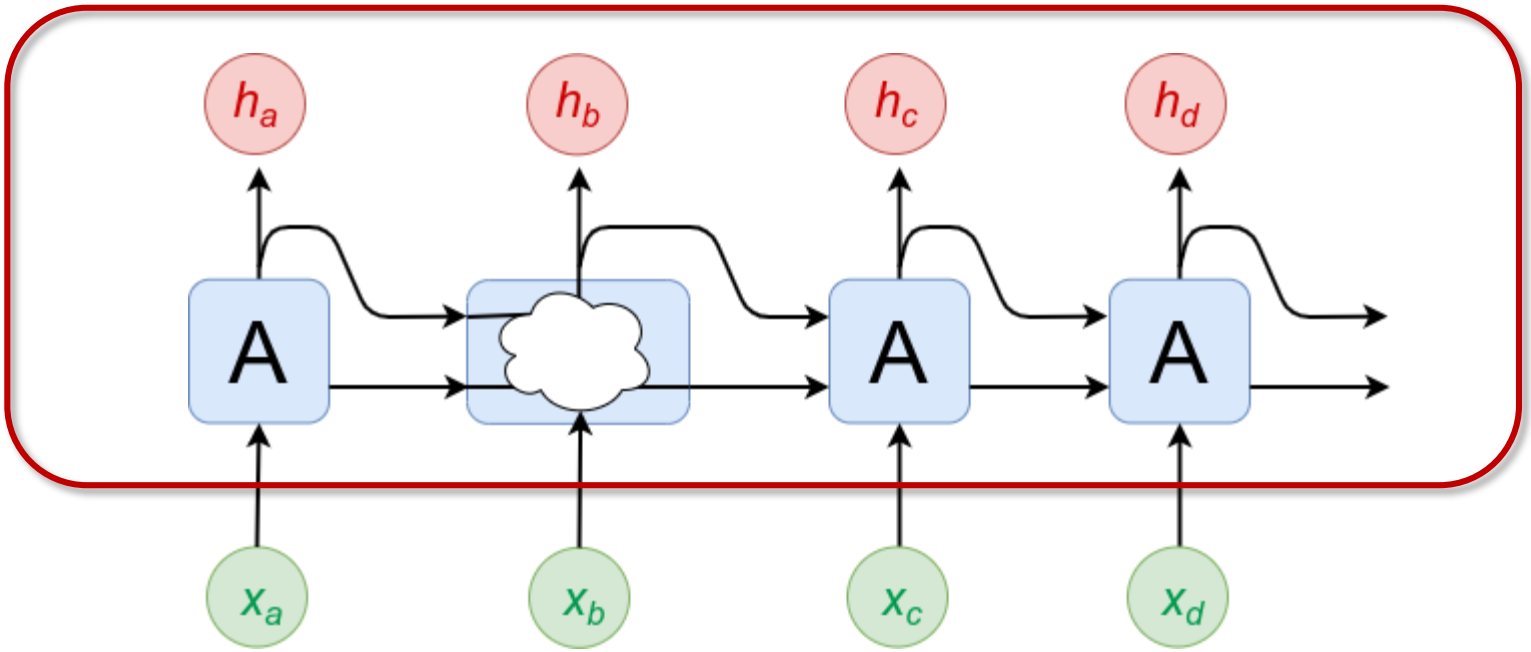
En pratique



Entrée : une matrice d'embeddings (une ligne = un mot)

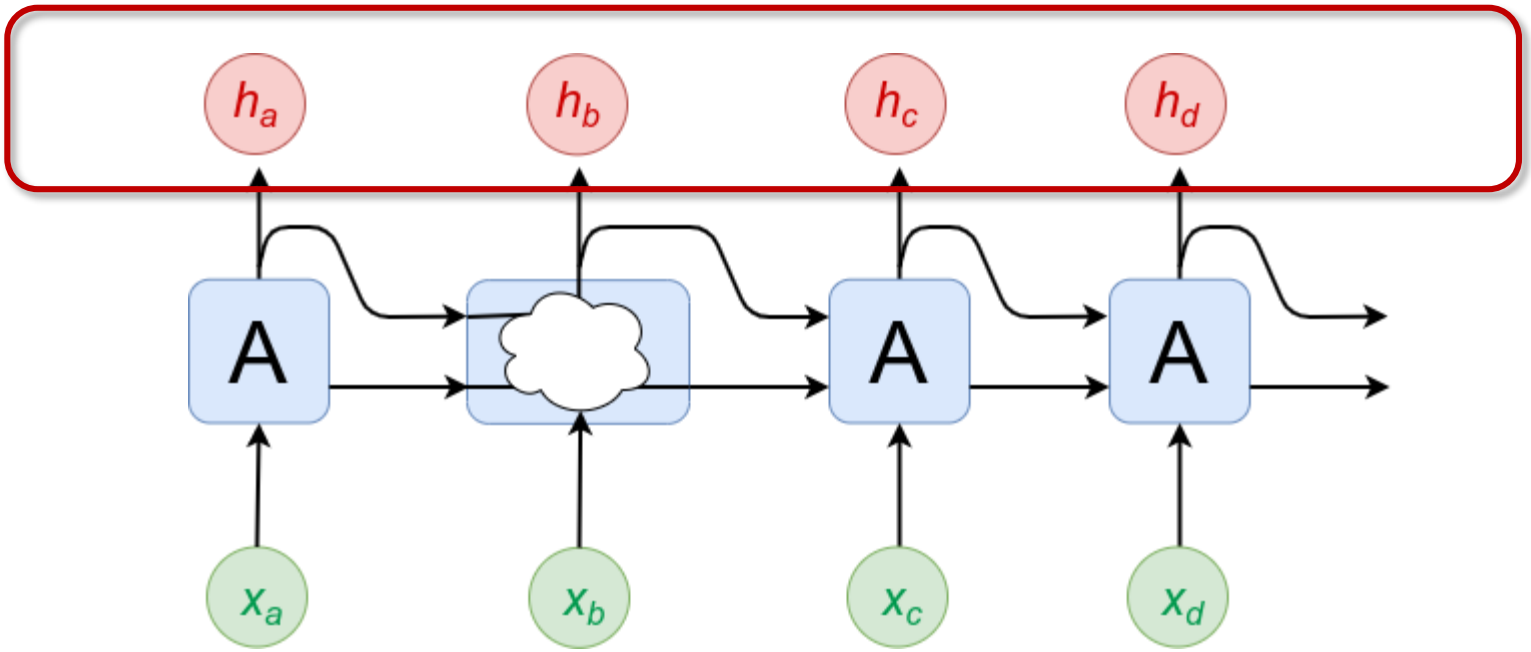
En pratique

Modèle : LSTM + couche de classification



En pratique

Évaluation : précision, rappel, F1-mesure (accuracy ?)



Ensuite

- ★ Convertir le LSTM en Bi-LSTM
- ★ Jouer avec les hyperparamètres
- ★★ Ajouter des embeddings pré entrainés (Word2vec, fasttext…)
- ★★★ Prétraitement ? (stemming…)
- ★ Changer de modèle de RNN (GRU ?)
- ★★ Ajouter des couches au modèle de classification
- ★★ Changer le format des données (IOB → IO, IOBES, etc.)
- ★★★★ Ajouter un CRF
- ★★★★ Ajouter une couche BERT

- ★★★★ Un modèle de représentation au niveau des caractères ? (CNN, RNN)
- ★★★ Optimiser les hyperparamètres avec Optuna