



TAL: traitement automatique de la langue

Manipulation des séquences

Vincent Guigue
UPMC - LIP6

Fonctionnement de l'UE

- ▶ TAL = mélange de plusieurs disciplines
 - 1/2 Traitement de la Langue (=métier): [Brigitte Grau & Anne-Laure Ligozat](#)
 - 1/2 Apprentissage Automatique (=outils): [Vincent Guigue](#)
- ▶ Evaluation:
 - 60% Projet = biblio + analyse d'une problématique TAL + manipulation des outils
 - 40% Examen = fonctionnement des méthodes d'apprentissage
- ▶ Positionnement:
 - Un peu de MAPSI/ARF...
 - Mais surtout de l'utilisation de modèles + campagne d'expériences (+professionalisante)



C'est quoi du texte?

- ▶ Une suite de lettres

l e c h a t e s t ...

- ▶ Une suite de mots

le chat est ...

- ▶ Un ensemble de mots

Dans l'ordre alphabétique

chat
est
le
...



Différentes tâches

- ▶ Classification de textes
 - Classification thématique
e.g. football, article scientifique, analyse politique
 - Classification de sentiments positif, négatif, neutre, agressif, ...
- ▶ Classification de mots
 - Grammaire = *Part-Of-Speech* (POS)
Nom, Nom propre, Déterminant, Verbe...
 - NER = *Named Entity Recognition*
détection des noms propres, travail sur les co-références
 - SRL = *Semantic Role Labeling* sujet, verbe, compléments...
- ▶ Question Answering, Extraction d'information



Applications

Part-of-Speech (POS) Tagging

- ▶ tags: ADJECTIVE, NOUN, PREPOSITION, VERB, ADVERB, ARTICLE...

Exemple: *Bob drank coffee at Starbucks*

⇒ Bob (NOUN) drank (VERB) coffee (NOUN) at (PREPOSITION) Starbucks (NOUN).

Named Entity Recognition (NER)

Sandra spent her holidays in Paris

Envoyer Clear

Sandra spent her holidays in **Paris**

- Potential tags:
- ORGANIZATION
 - LOCATION
 - PERSON

Applications (suite)

- ▶ Temporal tagger : reco + normalisation
- ▶ Parsing

Please enter a sentence to be parsed:

My dog also likes eating sausage.

Language:

Your query

My dog also likes eating sausage.

Tagging

My/PRP\$ dog/NN also/RB likes/VBZ eating/VBG sausage/NN ./.

Parse

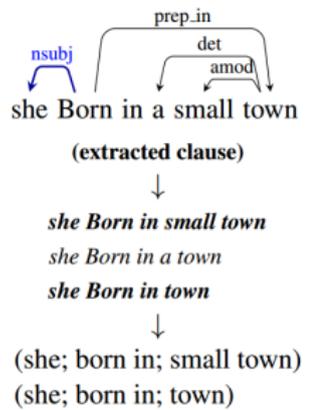
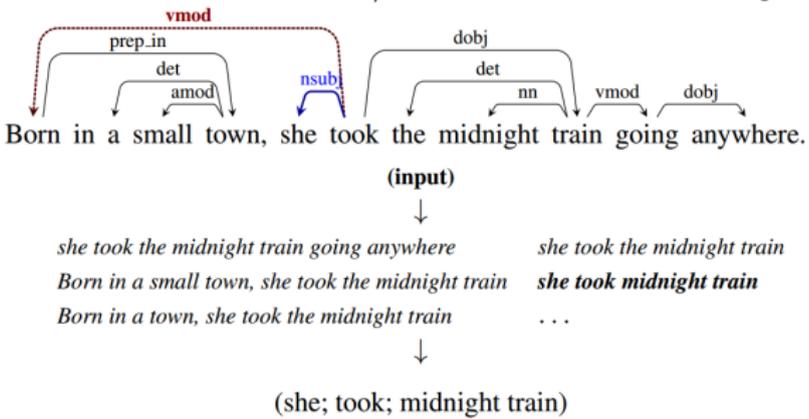
```
(ROOT
  (S
    (NP (PRP$ My) (NN dog))
    (ADVP (RB also))
    (VP (VBZ likes)
      (S
        (VP (VBG eating)
          (NP (NN sausage))))))
    (. .)))
```

crédit : CoreNLP



Applications (suite)

- ▶ Temporal tagger : reco + normalisation
- ▶ Parsing
- ▶ Information Extraction / Semantic Role Labeling :



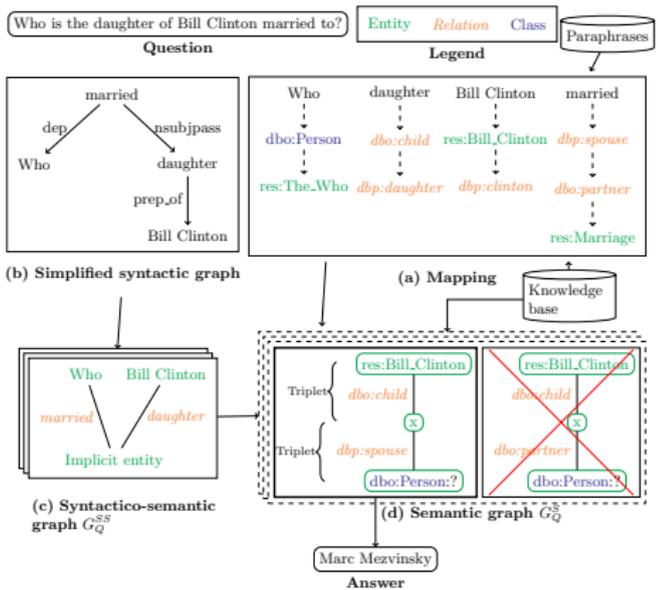
Crédit: Stanford NLP

- ▶ Question answering (QA)



Applications (suite)

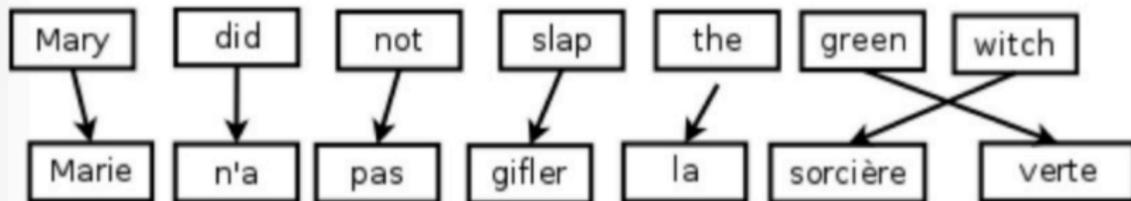
- ▶ Temporal tagger : reco + normalisation
- ▶ Parsing
- ▶ Information Extraction / Semantic Role Labeling :
- ▶ Question answering (QA)





Applications (suite)

- ▶ Traduction automatique
 - Aligner des mots
 - Générer une phrase intelligible / vraisemblable
- ▶ Historique (en évolution rapide)
 - traduction de mots
 - traduction de séquences
 - traduction de connaissances / signification





Modélisations du texte

Approche classique:

- ▶ Sac de mots (bag-of-words, BoW)

+ Avantages BoW

- plutôt simple, plutôt léger
- rapide (systèmes temps-réel, RI, indexation naturelle...)
- nb possibilités d'enrichissement (POS, codage du contexte, N-gram...)
- bien adapté pour la classification de documents
- Implémentations existantes efficaces `nlTK`, `sklearn`

– Inconvénient(s) BoW

- perte de la structure des phrases/documents
- ⇒ **Plusieurs tâches difficiles à attaquer**
- NER, POS tagging, SRL
 - Génération de textes

Mieux gérer les séquences

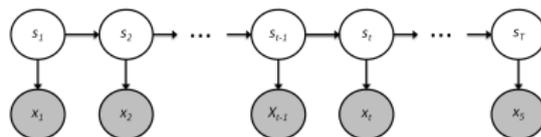
- 1 Enrichissement de la description vectorielle
 - N-grams,
 - Description du contexte des mots...
 - Usage type : amélioration des tâches de classification **au niveau document**

Mieux gérer les séquences

- 1 Enrichissement de la description vectorielle
 - N-grams,
 - Description du contexte des mots...
 - Usage type : amélioration des tâches de classification **au niveau document**
- 2 Approche par fenêtre glissante : prendre une décision à l'échelle intra-documentaire
 - Taille fixe \Rightarrow possibilité de description vectorielle
 - Classifieur sur une représentation locale du texte
 - Traitement du signal (AR, ARMA...)
 - Détection de pattern (*frequent itemsets*, règles d'association)

Mieux gérer les séquences

- 1 Enrichissement de la description vectorielle
 - N-grams,
 - Description du contexte des mots...
 - Usage type : amélioration des tâches de classification **au niveau document**
- 2 Approche par fenêtre glissante : prendre une décision à l'échelle intra-documentaire
 - Taille fixe \Rightarrow possibilité de description vectorielle
 - Classifieur sur une représentation locale du texte
 - Traitement du signal (AR, ARMA...)
 - Détection de pattern (*frequent itemsets*, règles d'association)
- 3 **Modèles séquentiels**
 - Hidden Markov Model (=Modèles de Markov Cachés)



- CRF (Conditional Random Fields) : approche discriminante



Historique des approches POS, SRL, NER

- ▶ Modélisation par **règle d'association** 80's
 - Quelles sont les cooccurrences fréquentes entre un POS et un item dans son contexte?
 - \Rightarrow Règles
- ▶ Modélisation **bayésienne**
 - Pour un POS i , modélisation de la distribution du contexte $p(\text{contexte}|\theta_i)$
 - Décision en MV: $\arg \max_i p(\text{contexte}|\theta_i)$
- ▶ Extension structurée (**Hidden Markov Model**) > 1985/90
 - *HMM taggers are fast and achieve precision/recall scores of about 93-95%*
- ▶ Vers une modélisation **discriminante (CRF)** > 2001
- ▶ **Recurrent Neural Network** (cf cours ARF, AS) > 2010



TAL / ML : beaucoup de choses en commun

- ▶ **Des financements liés** (et pas toujours glorieux) :
 - Conférences MUC / TREC (...)
 - RI, extraction d'info, classification de doc, sentiments, QA
 - Multiples domaines: général, médecine, brevet, judiciaire...
 - ⇒ Construction de bases, centralisation des résultats, échanges

Conference	Year	Text Source	Topic (Domain)
MUC-1	1987	Mil. reports	Fleet Operations
MUC-2	1989	Mil. reports	Fleet Operations
MUC-3	1991	News reports	Terrorist activities in Latin America
MUC-4	1992	News reports	Terrorist activities in Latin America
MUC-5	1993	News reports	Corporate Joint Ventures, Microelectronic production
MUC-6	1995	News reports	Negotiation of Labor Disputes and Corporate Management Succession
MUC-7	1997	News reports	Airplane crashes, and Rocket/Missile Launches

- NSA,
- Boites noires (loi sur le renseignement 2015)
- ▶ **Avancée ML tirée par le TAL** : HMM, CRF
- ▶ **De nombreux projets ambitieux**:
 - Google Knowledge Graph,
 - NELL: Never-Ending Language Learning (Tom Mitchell, CMU)

Formalisation séquentielle

Observations :		Le	chat	est	dans	le	salon
Etiquettes :		DET	NN	VBZ	...		

Le HMM est pertinent: il est basé sur

- ▶ les enchainements d'étiquettes,
- ▶ les probabilités d'observation



Notations

- ▶ La chaîne de Markov est toujours composée de:
 - d'une séquence d'**états** $S = (s_1, \dots, s_T)$
 - dont les valeurs sont tirées dans un ensemble fini $Q = (q_1, \dots, q_N)$
 - Le modèle est toujours défini par $\{\Pi, A\}$
 - $\pi_i = P(s_1 = q_i)$
 - $a_{ij} = p(s_{t+1} = q_j | s_t = q_i)$
- ▶ Les **observations** sont modélisées à partir des s_t
 - séquence d'observation: $X = (x_1, \dots, x_T)$
 - loi de probabilité: $b_j(t) = p(x_t | s_t = q_j)$
 - B peut être discrète ou continue
- ▶ MMC: $\lambda = \{\Pi, A, B\}$



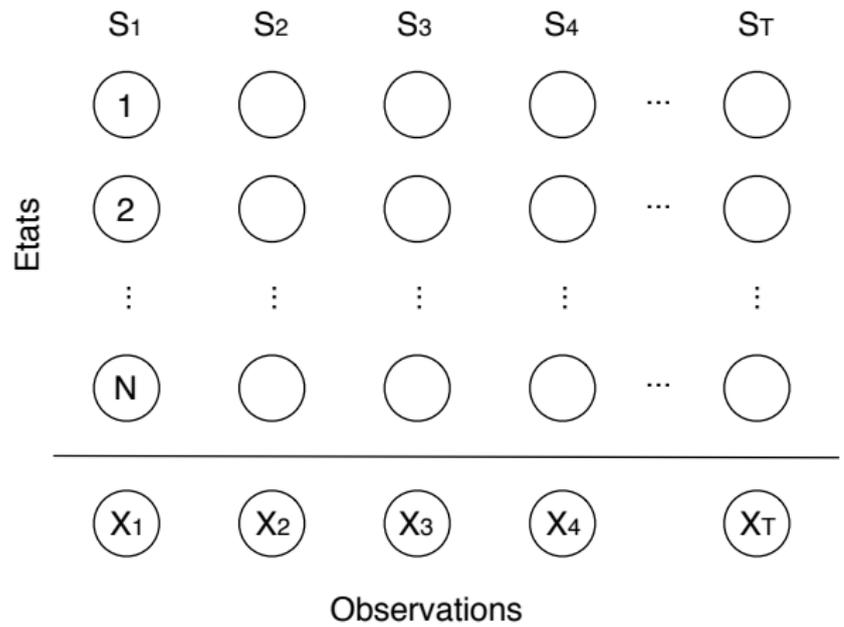
Ce que l'on manipule

- ▶ Séquence d'observations
 - $X = (x_1, \dots, x_T)$
- ▶ Séquence d'états (**cachée = manquante**)
 - $S = (s_1, \dots, s_T)$



Rappels sur la structure d'un MMC

Constitution d'un MMC:

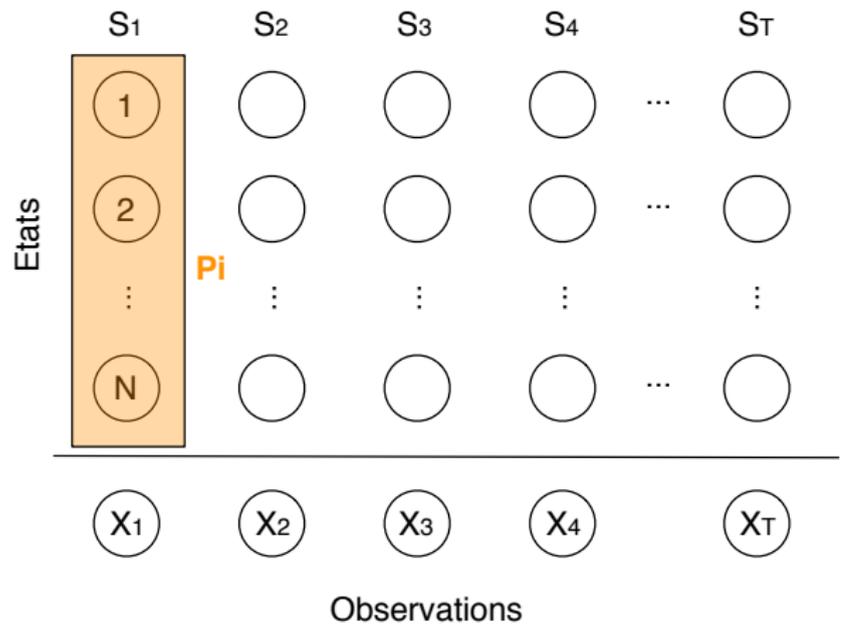


► Les états sont inconnus...



Rappels sur la structure d'un MMC

Constitution d'un MMC:

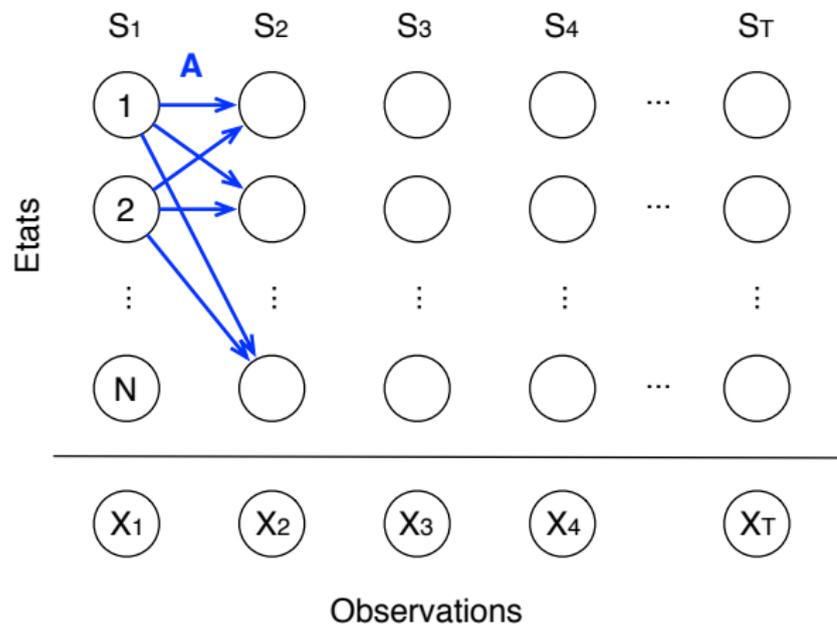


► Les états sont inconnus...



Rappels sur la structure d'un MMC

Constitution d'un MMC:

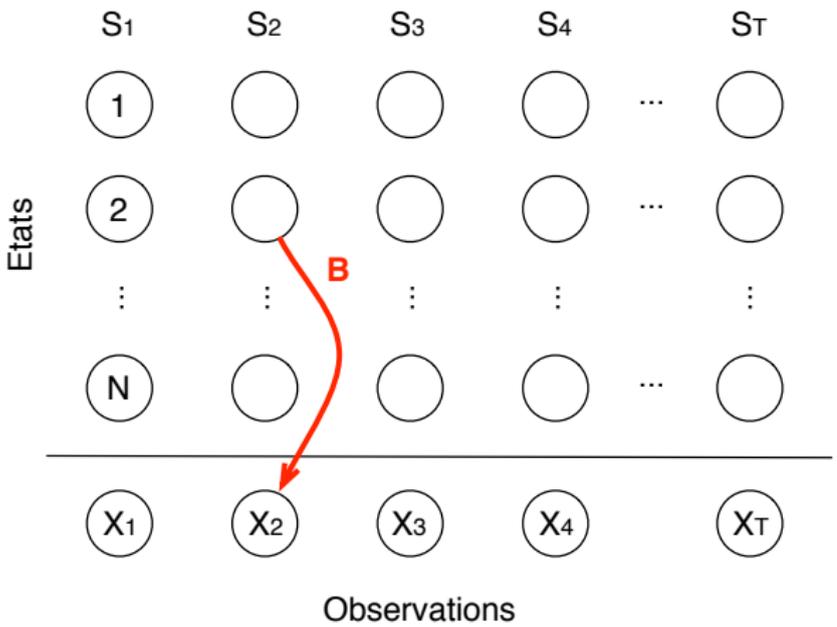


Hyp. Ordre 1:
chaque état ne dépend que du précédent

- ▶ Les états sont inconnus...
La combinatoire à envisager est problématique!

Rappels sur la structure d'un MMC

Constitution d'un MMC:



Hyp. Ordre 1:
chaque état ne dépend que du précédent

Chaque obs. ne dépend que de l'état courant

► Les états sont inconnus...

Les trois problèmes des MMC (Ferguson - Rabiner)

- ▶ **Evaluation:** λ donné, calcul de $p(x_1^T | \lambda)$
- ▶ **Décodage:** λ donné, quelle séquence d'états a généré les observations?

$$s_1^{T*} = \arg \max_{s_1^T} p(x_1^T, s_1^T | \lambda)$$

- ▶ **Apprentissage:** à partir d'une série d'observations, trouver λ^*

$$\lambda^* = \{\Pi^*, A^*, B^*\} = \arg \max_{s_1^T, \lambda} p(x_1^T, s_1^T | \lambda)$$



PB1: Algorithme *forward* (prog dynamique)

$$\alpha_t(i) = p(x_1^t, s_t = i | \lambda)$$

- ▶ Initialisation:

$$\alpha_{t=1}(i) = p(x_1^1, s_1 = i | \lambda) = \pi_i b_i(x_1)$$

- ▶ Itération:

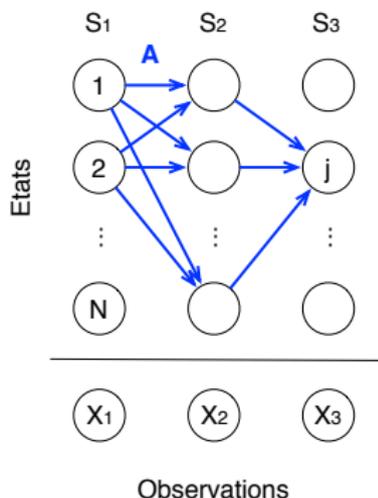
$$\alpha_t(j) = \left[\sum_{i=1}^N \alpha_{t-1}(i) a_{ij} \right] b_j(x_t)$$

- ▶ Terminaison:

$$p(x_1^T | \lambda) = \sum_{i=1}^N \alpha_T(i)$$

- ▶ Complexité linéaire en T

- Usuellement: $T \gg N$



PB2: Viterbi (récapitulatif)

$$\delta_t(i) = \max_{s_1^{t-1}} p(s_1^{t-1}, s_t = i, x_1^t | \lambda)$$

1 Initialisation

$$\begin{aligned} \delta_1(i) &= \pi_i b_i(x_1) \\ \Psi_1(i) &= 0 \end{aligned}$$

2 Récursion

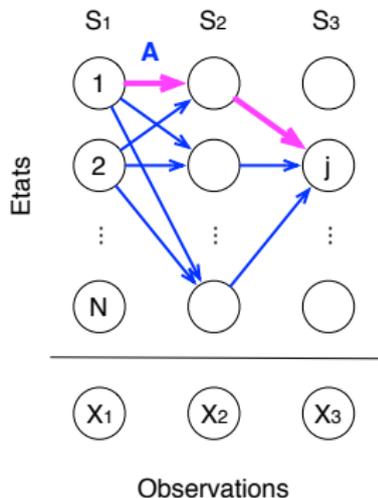
$$\begin{aligned} \delta_t(j) &= \left[\max_i \delta_{t-1}(i) a_{ij} \right] b_j(x_t) \\ \Psi_t(j) &= \arg \max_{i \in [1, N]} \delta_{t-1}(i) a_{ij} \end{aligned}$$

3 Terminaison

$$S^* = \max_i \delta_T(i)$$

4 Chemin

$$\begin{aligned} q_T^* &= \arg \max_i \delta_T(i) \\ q_t^* &= \Psi_{t+1}(q_{t+1}^*) \end{aligned}$$



PB3: Apprentissage des MMC

- ▶ **Version simplifiée** (*hard assignment*): type k -means
- ▶ Nous disposons de :
 - **Evaluation**: $p(x_1^T | \lambda)$
 - **Décodage**: $s_1^{T*} = \arg \max_{s_1^T} p(x_1^T | \lambda)$
- ▶ Proposition:

Data: Observations : X , Structure= N, K

Result: $\tilde{\Pi}^*, \tilde{A}^*, \tilde{B}^*$

Initialiser $\lambda_0 = \Pi^0, A^0, B^0$;

→ finement si possible;

$t = 0$;

while *convergence non atteinte* **do**

	$S_{t+1} = \text{decodage}(X, \lambda_t)$;
	$\lambda_{t+1}^* = \Pi^{t+1}, A^{t+1}, B^{t+1}$ obtenus par comptage des transitions ;
	$t = t + 1$;

end

Algorithm 1: Baum-Welch simplifié pour l'apprentissage d'un MMC

Vous avez déjà tous les éléments pour faire ça!



Apprentissage en contexte supervisé

Observations : | Le chat est dans le salon
Etiquettes : | DET NN VBZ ...

- ▶ Beaucoup plus simple (après la couteuse tâche d'étiquetage) :
- ▶ Matrices A, B, Π obtenues par comptage...
- ▶ Inférence = viterbi

Philosophie & limites:

Trouver l'étiquetage qui **maximise la vraisemblance de la séquence états-observations...**

... sous les hypothèses des HMM – indépendance des observations étant donnés les états, ordre 1 –

Intro

HMM

CRF

RNN



Modélisation discriminante vs générative

- ▶ Modélisation bayésienne = modèle génératif
 - Pour un POS i , modélisation de la distribution du contexte $p(\text{contexte}|\theta_i)$
 - Décision en MV: $\arg \max_i p(\text{contexte}|\theta_i)$
 - Génératif car on peut générer des contextes (même si ça n'a pas de sens dans cette application)
 - Très apprécié en TAL depuis les approches Naive Bayes (directement calculable en BD)



Modélisation discriminante vs générative

- ▶ Modélisation bayésienne = modèle génératif
 - Pour un POS i , modélisation de la distribution du contexte $p(\text{contexte}|\theta_i)$
 - Décision en MV: $\arg \max_i p(\text{contexte}|\theta_i)$
 - Génératif car on peut générer des contextes (même si ça n'a pas de sens dans cette application)
 - Très apprécié en TAL depuis les approches Naive Bayes (directement calculable en BD)
 - Bayésien + Séquence = HMM



Modélisation discriminante vs générative

- ▶ Modélisation bayésienne = modèle génératif
 - Pour un POS i , modélisation de la distribution du contexte $p(\text{contexte}|\theta_i)$
 - Décision en MV: $\arg \max_i p(\text{contexte}|\theta_i)$
 - Génératif car on peut générer des contextes (même si ça n'a pas de sens dans cette application)
 - Très apprécié en TAL depuis les approches Naive Bayes (directement calculable en BD)
 - Bayésien + Séquence = HMM
- ▶ Modélisation discriminante
 - Qu'est ce que distingue une classe d'une autre classe?
 - $p(y_i|\text{contexte})$: regression logistique
 - proba \Rightarrow scoring: fonction linéaire, minimisation d'un coût...



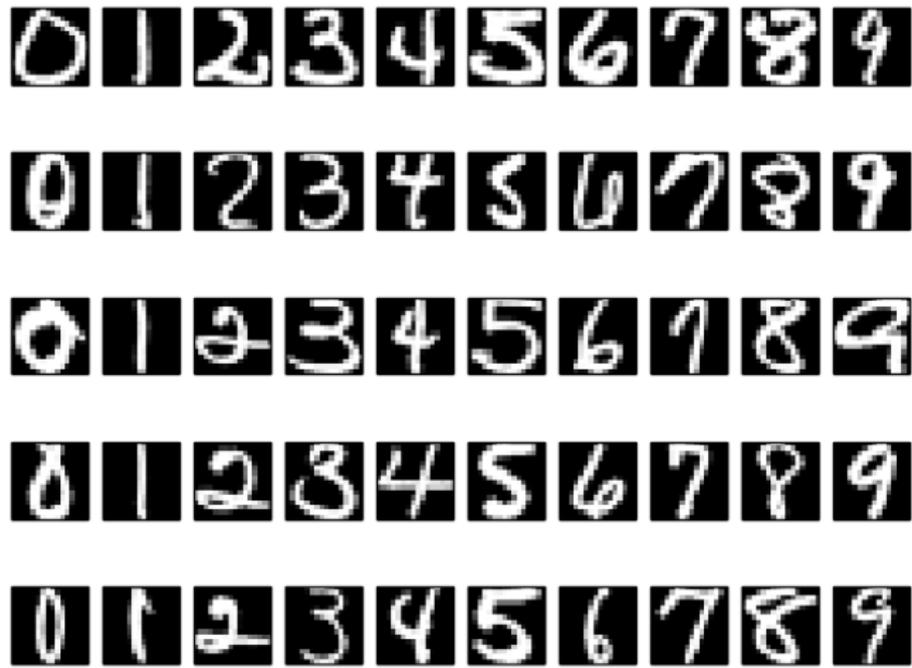
Modélisation discriminante vs générative

- ▶ Modélisation bayésienne = modèle génératif
 - Pour un POS i , modélisation de la distribution du contexte $p(\text{contexte}|\theta_i)$
 - Décision en MV: $\arg \max_i p(\text{contexte}|\theta_i)$
 - Génératif car on peut générer des contextes (même si ça n'a pas de sens dans cette application)
 - Très apprécié en TAL depuis les approches Naive Bayes (directement calculable en BD)
 - Bayésien + Séquence = HMM
- ▶ Modélisation discriminante
 - Qu'est ce que distingue une classe d'une autre classe?
 - $p(y_i|\text{contexte})$: regression logistique
 - proba \Rightarrow scoring: fonction linéaire, minimisation d'un coût...
 - Discriminant + séquence = CRF



Retour sur la regression logistique

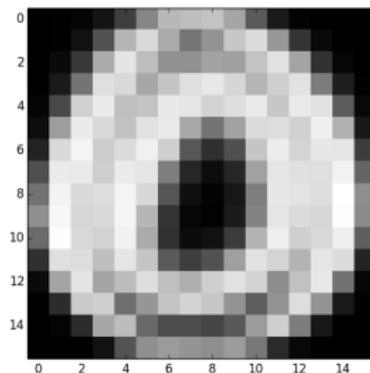
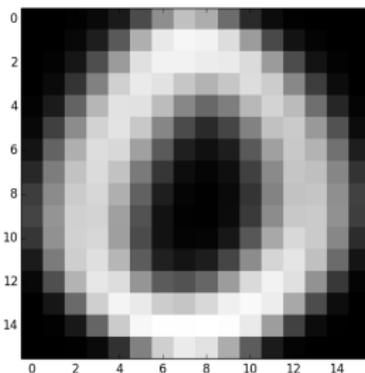
Le problème de reconnaissance des chiffres (USPS)



Retour sur la regression logistique

Le problème de reconnaissance des chiffres (USPS)

- ① Approche bayésienne générative :
 - modélisation de chaque classe (e.g. en gaussiennes (μ, σ))



Modélisation (μ, σ) de la classe 0 par MV = trouver ce qui caractérise un 0

- critère de décision MV :

$$y^* = \arg \max_y p(\mathbf{x}|y)$$

Quel modèle ressemble le plus à \mathbf{x} ?

Retour sur la regression logistique

Le problème de reconnaissance des chiffres (USPS)

- Approche bayésienne générative :
 - modélisation de chaque classe (e.g. en gaussiennes (μ, σ))
 - critère de décision MV :

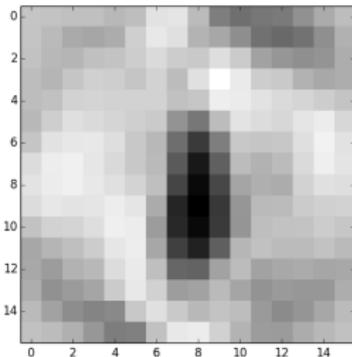
$$y^* = \arg \max_y p(\mathbf{x}|y)$$

Quel modèle ressemble le plus à \mathbf{x} ?

- Approche discriminante: modéliser directement $p(y|\mathbf{x})$
 - Construction d'un modèle minimisant les *a priori*:

$$p(Y = 1|\mathbf{x}) = \frac{1}{1 + \exp(-(\mathbf{x}\theta + \theta_0))}$$

Qu'est ce qui distingue un 0 des autres chiffres?





RegLog (2)

Processus bi-classe:

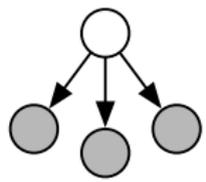
- ① $p(Y = 1|\mathbf{x}) = \frac{1}{1+\exp(-(\mathbf{x}\theta+\theta_0))}$
- ② $p(\mathbf{y}|\mathbf{x}) = P(Y = 1|\mathbf{x})^y \times [1 - P(Y = 1|\mathbf{x})]^{1-y}$, $y \in \{0, 1\}$
(Bernoulli)
- ③ (log) Vraisemblance : $L = \sum_{i=1}^N \log P(\mathbf{x}_i, y_i)$
- ④ Optimisation :
 - gradient : $\frac{\partial}{\partial \theta_j} L_{\log} = \sum_{i=1}^N x_{ij}(y_i - \frac{1}{1+\exp(\mathbf{x}\theta+\theta_0)})$
 - ... résolution analytique impossible...
 - montée de gradient $\theta_j \leftarrow \theta_j + \varepsilon \frac{\partial}{\partial \theta_j} L_{\log}$ ou BFGS

► Extension un-contre-tous

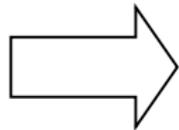


HMM \Rightarrow CRF

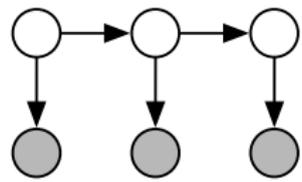
Introduction to CRF, [Sutton & McCallum](#)



Naive Bayes



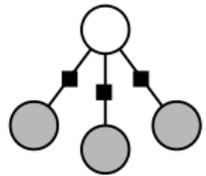
SEQUENCE



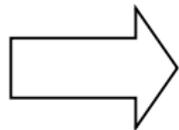
HMMs



CONDITIONAL



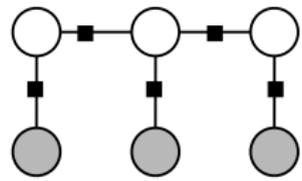
Logistic Regression



SEQUENCE



CONDITIONAL



Linear-chain CRFs



Modélisation CRF

- ▶ Séquence de mots $\mathbf{x} = \{x_1, \dots, x_T\}$
- ▶ Séquence d'étiquettes \mathbf{y} (=POS tag)

Estimation paramétrique des probabilités basées sur la famille exponentielle:

$$p(\mathbf{y}, \mathbf{x}) = \frac{1}{Z} \prod_t \Psi_t(y_t, y_{t-1}, \mathbf{x}_t),$$

$$\Psi_t(y_t, y_{t-1}, \mathbf{x}_t) = \exp(\sum_k \theta_k f_k(y_t, y_{t-1}, \mathbf{x}_t))$$

- ▶ Dépendances de $\Psi_t \Rightarrow$ forme du modèle HMM
- ▶ θ_k : paramètres à estimer (cf regression logistique)
- ▶ $f_k(y_t, y_{t-1}, \mathbf{x}_t)$: expression générique des caractéristiques
(détails plus loin)

CRF = généralisation des HMM

Cas général (slide précédent):

$$p(\mathbf{y}, \mathbf{x}) = \frac{1}{Z} \prod_t \exp \left(\sum_k \theta_k f_k(y_t, y_{t-1}, \mathbf{x}_t) \right)$$

Cas particulier : $f_k =$ existence de (y_t, y_{t-1}) ou (y_t, \mathbf{x}_t)

$$p(\mathbf{y}, \mathbf{x}) = \frac{1}{Z} \prod_t \exp \left(\sum_{i,j \in \mathcal{S}} \theta_{i,j} \mathbf{1}_{y_t=i \& y_{t-1}=j} + \sum_{i \in \mathcal{S}, o \in \mathcal{O}} \mu_{o,i} \mathbf{1}_{y_t=i \& x_t=o} \right)$$

Avec :

- ▶ $\theta_{i,j} = \log p(y_t = i | y_{t-1} = j)$
- ▶ $\mu_{o,i} = \log p(x = o | y_t = i)$
- ▶ $Z = 1$

⇒ Dans ce cas, les caractéristiques sont binaires (1/0)



CRF : passage aux probas conditionnelles

$$p(\mathbf{y}, \mathbf{x}) = \frac{1}{Z} \prod_t \exp \left(\sum_k \theta_k f_k(y_t, y_{t-1}, \mathbf{x}_t) \right)$$

⇒

$$\begin{aligned} p(\mathbf{y}|\mathbf{x}) &= \frac{\prod_t \exp(\sum_k \theta_k f_k(y_t, y_{t-1}, \mathbf{x}_t))}{\sum_{y'} \prod_t \exp(\sum_k \theta_k f_k(y'_t, y'_{t-1}, \mathbf{x}_t))} \\ &= \frac{1}{Z(\mathbf{x})} \prod_t \exp \left(\sum_k \theta_k f_k(y_t, y_{t-1}, \mathbf{x}_t) \right) \end{aligned}$$



Apprentissage CRF

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \prod_t \exp \left(\sum_k \theta_k f_k(y_t, y_{t-1}, \mathbf{x}_t) \right)$$

Comme pour la regression logistique:

$$\mathcal{L}_{cond} = \sum_n \log p(\mathbf{y}^n | \mathbf{x}^n) =$$

$$\sum_n \sum_t \sum_k \theta_k f_k(\mathbf{x}^n, y_t^n, y_{t-1}^n) - \sum_n \log(Z(\mathbf{x}^n))$$

Comment optimiser?

Apprentissage CRF

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \prod_t \exp \left(\sum_k \theta_k f_k(y_t, y_{t-1}, \mathbf{x}_t) \right)$$

Comme pour la regression logistique:

$$\mathcal{L}_{cond} = \sum_n \log p(\mathbf{y}^n | \mathbf{x}^n) =$$

$$\sum_n \sum_t \sum_k \theta_k f_k(\mathbf{x}^n, y_t^n, y_{t-1}^n) - \sum_n \log(Z(\mathbf{x}^n))$$

Comment optimiser?

Montée de gradient $\frac{\partial \mathcal{L}}{\partial \theta_k}$

$$\theta_k \leftarrow \theta_k + \sum_{n,t} f_k(\mathbf{x}^n, y_t^n, y_{t-1}^n) - \sum_{n,t} \sum_{y'_t, y'_{t-1}} f_k(\mathbf{x}^n, y'_t, y'_{t-1}) p(y'_t, y'_{t-1} | \mathbf{x}^n)$$

Calcul exact possible $\mathcal{O}(TM^2N)$, solutions approchées plus rapides

- M : nb labels, T : lg chaîne, N : nb chaînes

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \prod_t \exp \left(\sum_k \theta_k f_k(y_t, y_{t-1}, \mathbf{x}_t) \right)$$

Comme pour la regression logistique:

$$\mathcal{L}_{cond} = \sum_n \log p(\mathbf{y}^n | \mathbf{x}^n) =$$

$$\sum_n \sum_t \sum_k \theta_k f_k(\mathbf{x}^n, y_t^n, y_{t-1}^n) - \sum_n \log(Z(\mathbf{x}^n))$$

Comment optimiser?

- ▶ La difficulté réside essentiellement dans le facteur de normalisation



Régularisation

- ▶ Que pensez-vous de la complexité du modèle?
- ▶ Comment la limiter?

$$\mathcal{L}_{cond} = \sum_{n,k,t} \theta_k f_k(\mathbf{x}^n, y_t^n, y_{t-1}^n) - \sum_n \log(Z(\mathbf{x}_n))$$

⇒

$$\mathcal{L}_{cond} = \sum_{n,k,t} \theta_k f_k(\mathbf{x}^n, y_t^n, y_{t-1}^n) - \sum_n \log(Z(\mathbf{x}_n)) + \frac{1}{2\sigma^2} \|\theta\|^2$$

$$\mathcal{L}_{cond} = \sum_{n,k,t} \theta_k f_k(\mathbf{x}^n, y_t^n, y_{t-1}^n) - \sum_n \log(Z(\mathbf{x}_n)) + \alpha \sum_k |\theta_k|$$



Et les f_j alors???

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp \left\{ \sum_k \sum_{t=1}^T \theta_k f_k(\mathbf{x}, y_t, y_{t-1}) \right\}$$

- ▶ Par défaut, les f_k sont des *features* et **ne sont pas apprises**
- ▶ Exemples:
 - $f_1(\mathbf{x}, y_t, y_{t-1}) = 1$ if $y_t = \text{ADVERB}$ and the word ends in "-ly"; 0 otherwise.
 - If the weight θ_1 associated with this feature is large and positive, then this feature is essentially saying that we prefer labelings where words ending in -ly get labeled as ADVERB.
 - $f_2(\mathbf{x}, y_t, y_{t-1}) = 1$ si $t=1$, $y_t = \text{VERB}$, and the sentence ends in a question mark; 0 otherwise.
 - $f_3(\mathbf{x}, y_t, y_{t-1}) = 1$ if $y_{t-1} = \text{ADJECTIVE}$ and $y_t = \text{NOUN}$; 0 otherwise.
- ▶ Il est possible d'apprendre des features sur des données



Et les f_j alors???

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp \left\{ \sum_k \sum_{t=1}^T \theta_k f_k(\mathbf{x}, y_t, y_{t-1}) \right\}$$

- ▶ Par défaut, les f_k sont des *features* et **ne sont pas apprises**
- ▶ Exemples:
 - $f_1(\mathbf{x}, y_t, y_{t-1}) = 1$ if $y_t = \text{ADVERB}$ and the word ends in "-ly"; 0 otherwise.
 - $f_2(\mathbf{x}, y_t, y_{t-1}) = 1$ si $t=1$, $y_t = \text{VERB}$, and the sentence ends in a question mark; 0 otherwise.
 - Again, if the weight θ_2 associated with this feature is large and positive, then labelings that assign VERB to the first word in a question (e.g., Is this a sentence beginning with a verb??) are preferred.
 - $f_3(\mathbf{x}, y_t, y_{t-1}) = 1$ if $y_{t-1} = \text{ADJECTIVE}$ and $y_t = \text{NOUN}$; 0 otherwise.
- ▶ Il est possible d'apprendre des features sur des données



Et les f_j alors???

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp \left\{ \sum_k \sum_{t=1}^T \theta_k f_k(\mathbf{x}, y_t, y_{t-1}) \right\}$$

- ▶ Par défaut, les f_k sont des *features* et **ne sont pas apprises**
- ▶ Exemples:
 - $f_1(\mathbf{x}, y_t, y_{t-1}) = 1$ if $y_t = \text{ADVERB}$ and the word ends in "-ly"; 0 otherwise.
 - $f_2(\mathbf{x}, y_t, y_{t-1}) = 1$ si $t=1$, $y_t = \text{VERB}$, and the sentence ends in a **question mark**; 0 otherwise.
 - $f_3(\mathbf{x}, y_t, y_{t-1}) = 1$ if $y_{t-1} = \text{ADJECTIVE}$ and $y_t = \text{NOUN}$; 0 otherwise.
 - Again, a positive weight for this feature means that adjectives tend to be followed by nouns.
- ▶ Il est possible d'apprendre des features sur des données



Feature engineering

- ▶ **Label-observation features** : les f_k qui s'expriment $f_k(\mathbf{x}, y_t) = \mathbf{1}_{y_t=c} q_k(\mathbf{x})$ sont plus facile à calculer (ils sont calculés une fois pour toutes)
 - e.g. : si $\mathbf{x} = \text{mot}_i$, \mathbf{x} termine par *-ing*, \mathbf{x} a une majuscule... alors 1 sinon 0
- ▶ **Node-obs feature** : même si c'est moins précis, essayer de ne pas mélanger les références sur les observations et sur les transitions.
 - $f_k(\mathbf{x}, y_t, y_{t-1}) = q_k(\mathbf{x}) \mathbf{1}_{y_t=c} \mathbf{1}_{y_{t-1}=c'} \Rightarrow$
 $f_k(x_t, y_t) = q_k(\mathbf{x}) \mathbf{1}_{y_t=c}$, & $f_{k+1}(y_t, y_{t-1}) = \mathbf{1}_{y_t=c} \mathbf{1}_{y_{t-1}=c'}$
- ▶ **Boundary labels**



Feature engineering (2)

- ▶ **Unsupported features** : générée automatiquement à partir des observations (e.g. *with* n'est pas un nom de ville)... Mais pas très pertinente.
 - Utiliser ces features pour désambigüiser les erreurs dans la base d'apprentissage
- ▶ **Feature induction**
- ▶ **Features from different time steps**
- ▶ **Redundant features**
- ▶ **Complexe features** = model outputs



Processus:

- 1 Définir des caractéristiques,
- 2 Apprendre les paramètres du modèles (θ)
- 3 Classifier les nouvelles phrases (=inférence)

$$\mathbf{y}^* = \arg \max_{\mathbf{y}} p(\mathbf{y}|\mathbf{x})$$

⇒ Solution très proche de l'algorithme de Viterbi



Inférence (2)

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \prod_t \Psi_t(y_t, y_{t-1}, \mathbf{x}_t),$$

$$\Psi_t(y_t, y_{t-1}, \mathbf{x}_t) = \exp \left(\sum_k \theta_k f_k(y_t, y_{t-1}, \mathbf{x}_t) \right)$$

- 1 Passerelle avec les HMM:

$$\Psi_t(i, j, \mathbf{x}) = p(y_t = j | y_{t-1} = i) p(x_t = \mathbf{x} | y_t = j)$$

$$\alpha_t(j) = \sum_i \Psi_t(i, j, \mathbf{x}) \alpha_{t-1}(i), \quad \beta_t(i) = \sum_j \Psi_t(i, j, \mathbf{x}) \beta_{t+1}(j)$$

$$\delta_t(j) = \max_i \Psi_t(i, j, \mathbf{x}) \delta_{t-1}(i)$$



Inférence (2)

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \prod_t \Psi_t(y_t, y_{t-1}, \mathbf{x}_t),$$

$$\Psi_t(y_t, y_{t-1}, \mathbf{x}_t) = \exp \left(\sum_k \theta_k f_k(y_t, y_{t-1}, \mathbf{x}_t) \right)$$

- 1 Passerelle avec les HMM:

$$\Psi_t(i, j, \mathbf{x}) = p(y_t = j | y_{t-1} = i) p(\mathbf{x}_t = \mathbf{x} | y_t = j)$$

$$\alpha_t(j) = \sum_i \Psi_t(i, j, \mathbf{x}) \alpha_{t-1}(i), \quad \beta_t(i) = \sum_j \Psi_t(i, j, \mathbf{x}) \beta_{t+1}(j)$$

$$\delta_t(j) = \max_i \Psi_t(i, j, \mathbf{x}) \delta_{t-1}(i)$$

- 2 Les définitions restent valables avec les CRF (cf Sutton, McCallum), avec:

$$Z(\mathbf{x}) = \sum_i \alpha_T(i) = \beta_0(y_0)$$



Applications



- ▶ Analyse des phrases: analyse morpho-syntaxique
- ▶ NER: named entity recognition

...	Mister	George	W.	Bush	arrived	in	Rome	together	with	...
...	O	name	name	name	O	O	city	O	O	...

- ▶ Passage en 2D... Analyse des images



Applications

- ▶ Analyse des phrases: analyse morpho-syntaxique
- ▶ NER: named entity recognition

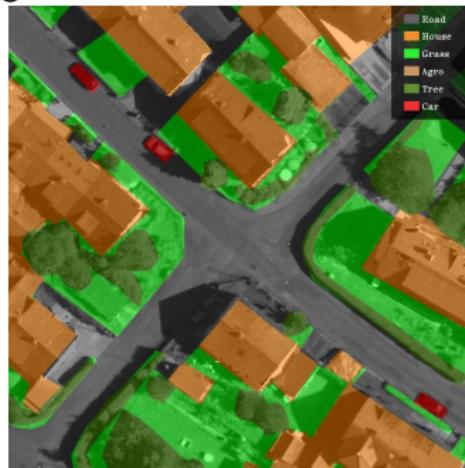
...	Mister	George	W.	Bush	arrived	in	Rome	together	with	...
...	O	name	name	name	O	O	city	O	O	...

- ▶ Passage en 2D... Analyse des images

- Détection des contours
- Classification d'objets

features =

- cohésion de l'espace
- enchainements usuels/impossibles



crédit: DGM lib



Intro

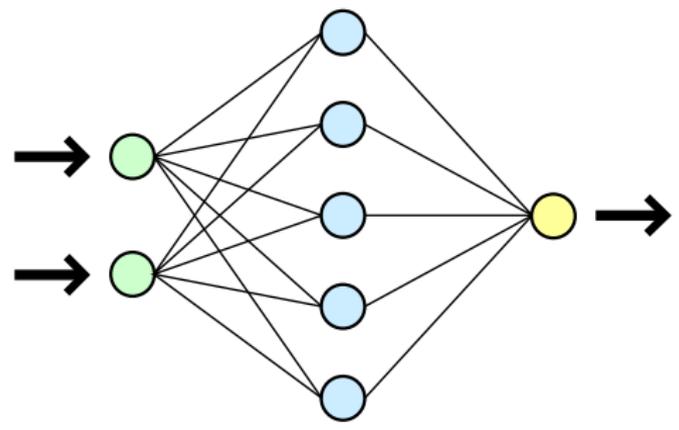
HMM

CRF

RNN

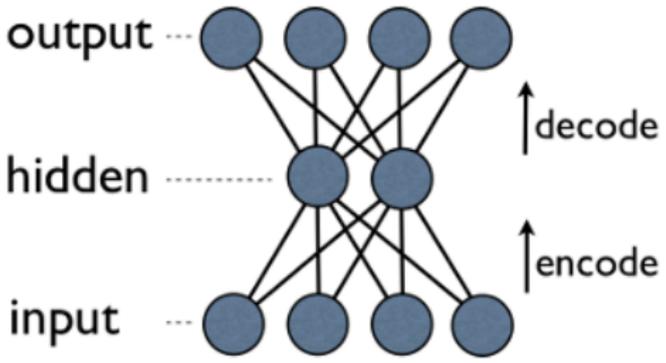
Réseaux de neurones

Architecture non linéaire pour la classification ou la régression



LP Réseaux de neurones

- ▶ Codage BoW, application en classification...
- ▶ Variante avec auto-encodeurs



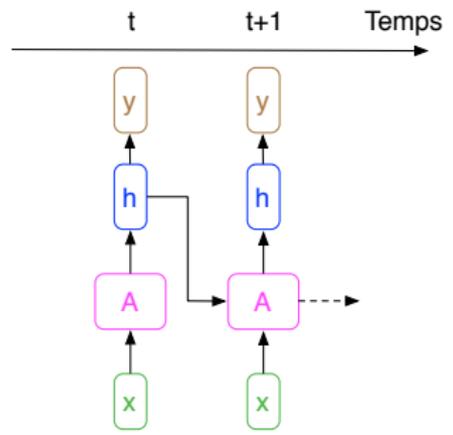


Réseaux de neurones récurrents

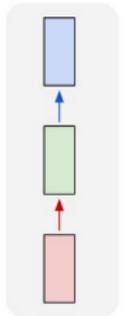


Un réseau pour suivre le texte:

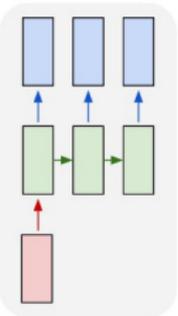
- ▶ Apprentissage de dépendances longues...
- ▶ Différentes architectures pour différentes applications



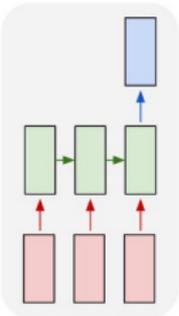
one to one



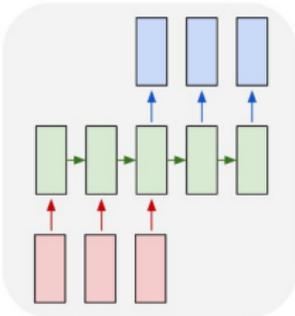
one to many



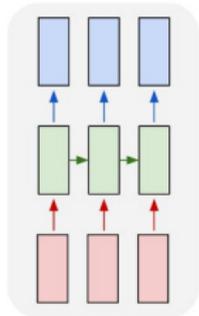
many to one



many to many

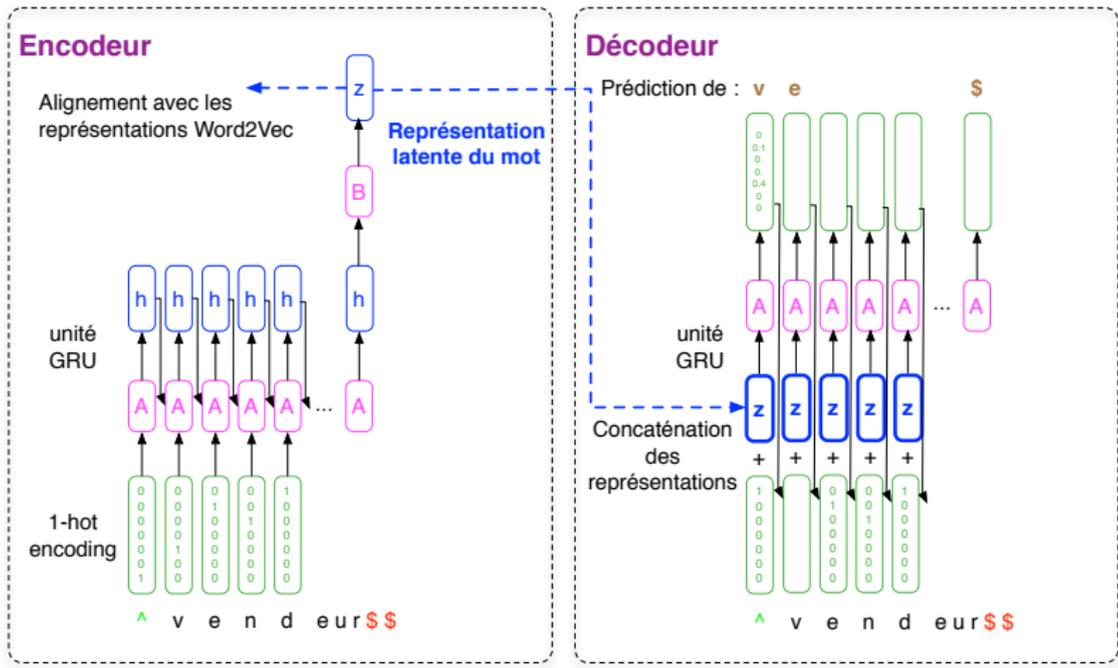


many to many





RNN génératifs





Exemples de générations (Shakespeare)

PANDARUS:

Alas, I think he shall be come approached and the day
When little srain would be attain'd into being never fed,
And who is but a chain and subjects of his death,
I should not sleep.

Second Senator:

They are away this miseries, produced upon my soul,
Breaking and strongly should be buried, when I perish
The earth and thoughts of many states.



Exemples de générations (wikipedia)

Naturalism and decision for the majority of Arab countries' cap
 by the Irish language by [[John Clair]], [[An Imperial Japanese
 with Guangzham's sovereignty. His generals were the powerful rul
 in the [[Protestant Immineners]], which could be said to be dire
 Communication, which followed a ceremony and set inspired prison
 emperor travelled back to [[Antioch, Perth, October 25|21]] to n
 of Costa Rica, unsuccessful fashioned the [[Thrales]]



Karpathy's blog



Exemples de générations (wikipedia)

Proof. Omitted. □

Lemma 0.1. Let \mathcal{C} be a set of the construction.
 Let \mathcal{C} be a gerber covering. Let \mathcal{F} be a quasi-coherent sheaves of \mathcal{O} -modules. We have to show that

$$\mathcal{O}_{\mathcal{O}_X} = \mathcal{O}_X(\mathcal{L})$$

Proof. This is an algebraic space with the composition of sheaves \mathcal{F} on $X_{\text{étale}}$ we have

$$\mathcal{O}_X(\mathcal{F}) = \{ \text{morph}_1 \times_{\mathcal{O}_X} (\mathcal{G}, \mathcal{F}) \}$$

where \mathcal{G} defines an isomorphism $\mathcal{F} \rightarrow \mathcal{F}$ of \mathcal{O} -modules. □

Lemma 0.2. This is an integer Z is injective.

Proof. See Spaces, Lemma ?? □

Lemma 0.3. Let S be a scheme. Let X be a scheme and X is an affine open covering. Let $\mathcal{U} \subset X$ be a canonical and locally of finite type. Let X be a scheme. Let X be a scheme which is equal to the formal complex.

The following to the construction of the lemma follows.

Let X be a scheme. Let X be a scheme covering. Let

$$b : X \rightarrow Y' \rightarrow Y \rightarrow Y \rightarrow Y' \times_X Y \rightarrow X.$$

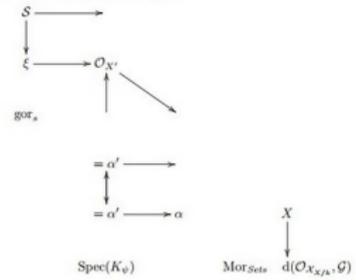
be a morphism of algebraic spaces over S and Y .

Proof. Let X be a nonzero scheme of X . Let X be an algebraic space. Let \mathcal{F} be a quasi-coherent sheaf of \mathcal{O}_X -modules. The following are equivalent

- (1) \mathcal{F} is an algebraic space over S .
- (2) If X is an affine open covering.

Consider a common structure on X and X the functor $\mathcal{O}_X(U)$ which is locally of finite type. □

This since $\mathcal{F} \in \mathcal{F}$ and $x \in \mathcal{G}$ the diagram



is a limit. Then \mathcal{G} is a finite type and assume S is a flat and \mathcal{F} and \mathcal{G} is a finite type f_* . This is of finite type diagrams, and

- the composition of \mathcal{G} is a regular sequence,
- $\mathcal{O}_{X'}$ is a sheaf of rings.

Proof. We have see that $X = \text{Spec}(R)$ and \mathcal{F} is a finite type representable by algebraic space. The property \mathcal{F} is a finite morphism of algebraic stacks. Then the cohomology of X is an open neighbourhood of U . □

Proof. This is clear that \mathcal{G} is a finite presentation, see Lemmas ??.
 A reduced above we conclude that U is an open covering of \mathcal{C} . The functor \mathcal{F} is a "field"

$$\mathcal{O}_{X,x} \rightarrow \mathcal{F}_x \rightarrow \mathcal{O}_{X_{\text{étale}}} \rightarrow \mathcal{O}_{X'} \rightarrow \mathcal{O}_{X'}(\mathcal{O}_{X'}^{\vee})$$

is an isomorphism of covering of $\mathcal{O}_{X'}$. If \mathcal{F} is the unique element of \mathcal{F} such that X is an isomorphism.

The property \mathcal{F} is a disjoint union of Proposition ?? and we can filter set of presentations of a scheme \mathcal{O}_X -algebra with \mathcal{F} are opens of finite type over S . If \mathcal{F} is a scheme theoretic image points. □

If \mathcal{F} is a finite direct sum \mathcal{O}_{X_i} is a closed immersion, see Lemma ?? . This is a sequence of \mathcal{F} is a similar morphism.



Exemples de générations (linux code)

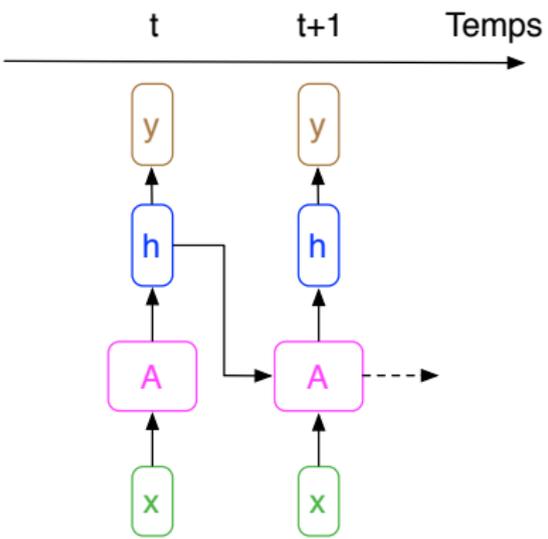
```

*
* Increment the size file of the new incorrect UI_FILTER group
* of the size generatively.
*/
static int indicate_policy(void)
{
    int error;
    if (fd == MARN_EPT) {
        /*
         * The kernel blank will coeld it to userspace.
         */
        if (ss->segment < mem_total)
            unblock_graph_and_set_blocked();
        else
            ret = 1;
        goto bail;
    }
}

```

Approches par RNN

Karpathy ⇒ Générer du texte = Utiliser des RNN



Neurones exotiques / gestion de la mémoire:

- ▶ GRU
- ▶ LSTM

<http://colah.github.io/posts/2015-08-Understanding-LSTMs/>

Inférence *beam search*: maximiser la probabilité de la séquence.



Andrej Karpathy blog, 2015

The Unreasonable Effectiveness of Recurrent Neural Networks

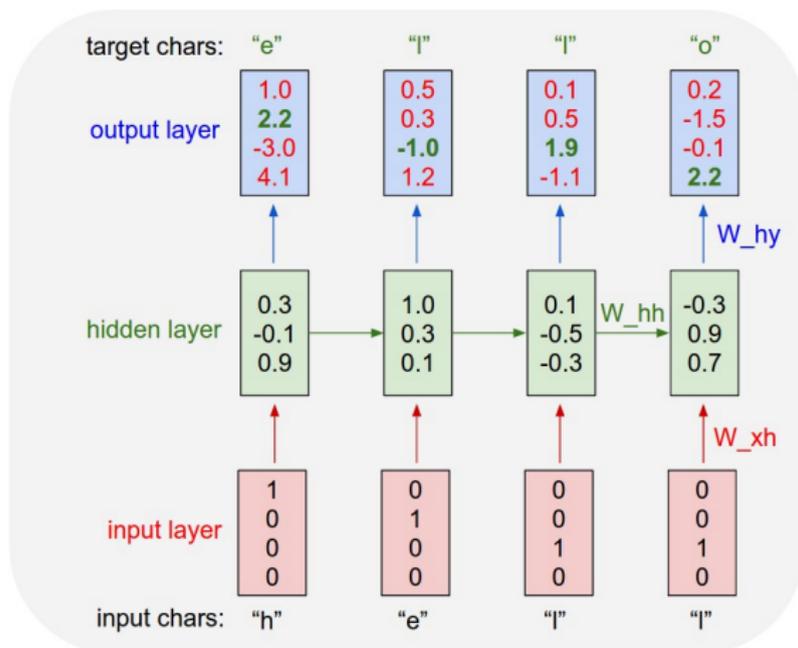


Lipton, Vikram, McAuley, arXiv, 2016

Generative Concatenative Nets Jointly Learn to Write and Classify Reviews



Approches par RNN



Andrej Karpathy blog, 2015

The Unreasonable Effectiveness of Recurrent Neural Networks

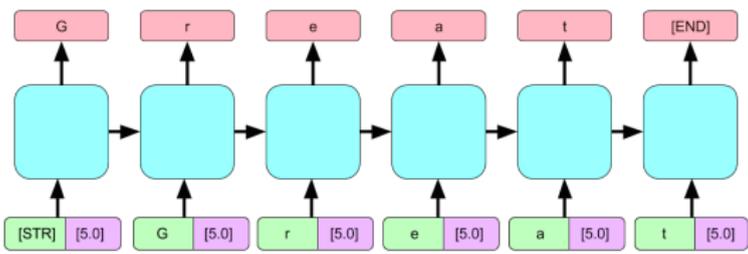


Iipton, Vikram, McAuley, arXiv, 2016

Generative Concatenative Nets Jointly Learn to Write and Classify Reviews



RNN & modélisation du contexte

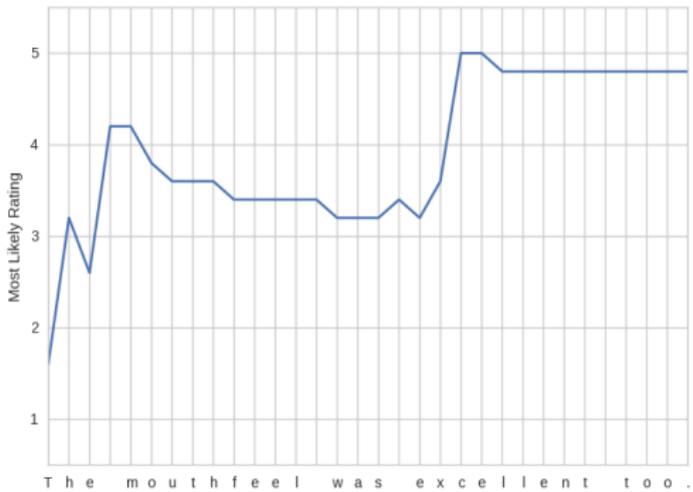


- ▶ Concatenation: input + contexte
- ▶ Inférence = test (combinatoire) des contextes + max de vraisemblance

Démo: <http://deepx.ucsd.edu/>

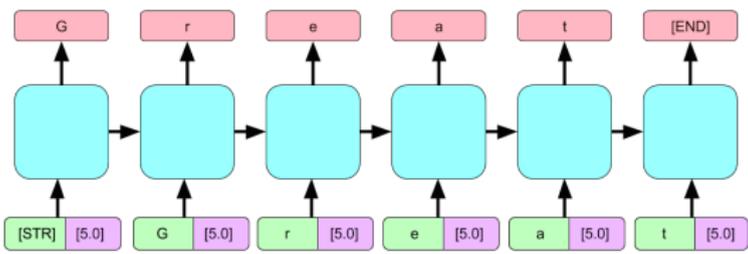


Lipton, Vikram, McAuley, arXiv, 2016
Gen. Concatenative Nets Jointly Learn to Write and Classify Reviews





RNN & modélisation du contexte

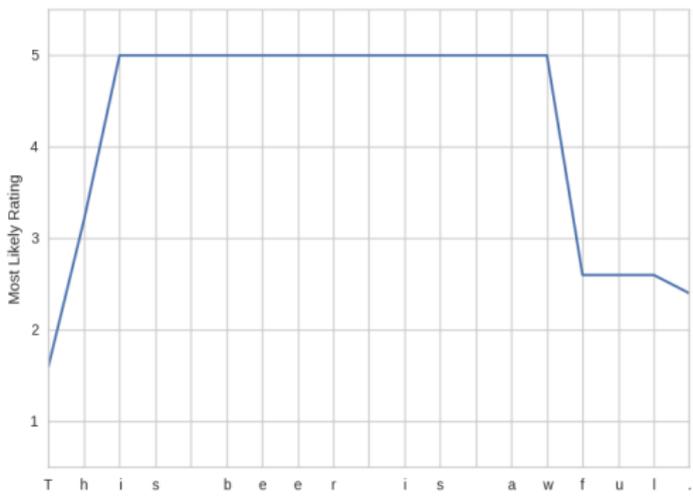


- ▶ Concatenation: input + contexte
- ▶ Inférence = test (combinatoire) des contextes + max de vraisemblance

Démo: <http://deepx.ucsd.edu/>

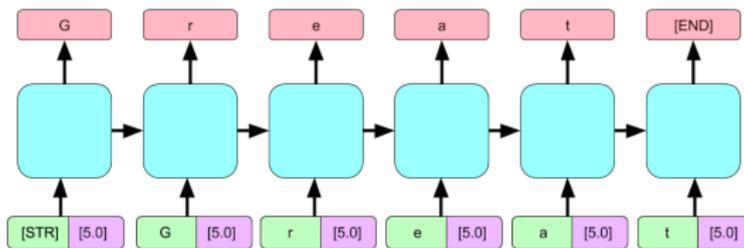


Lipton, Vikram, McAuley, arXiv, 2016
Gen. Concatenative Nets Jointly Learn to Write and Classify Reviews





RNN & modélisation du contexte

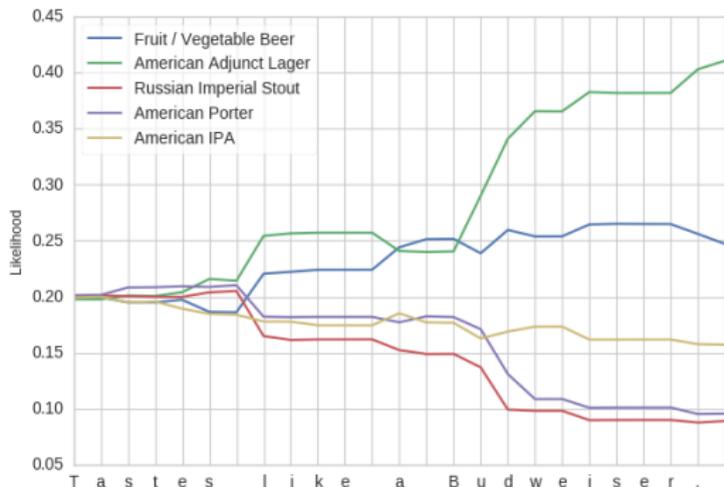


- ▶ Concatenation: input + contexte
- ▶ Inférence = test (combinatoire) des contextes + max de vraisemblance

Démo: <http://deepx.ucsd.edu/>

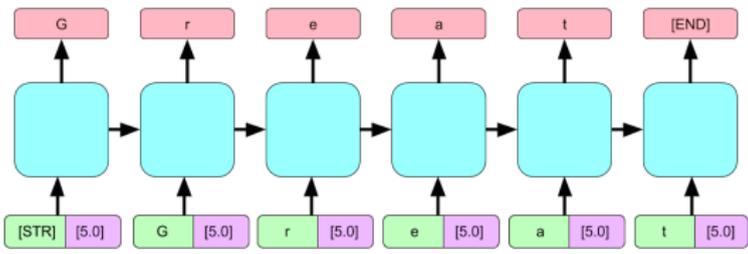


Lipton, Vikram, McAuley, arXiv, 2016
Gen. Concatenative Nets Jointly Learn to Write and Classify Reviews





RNN & modélisation du contexte

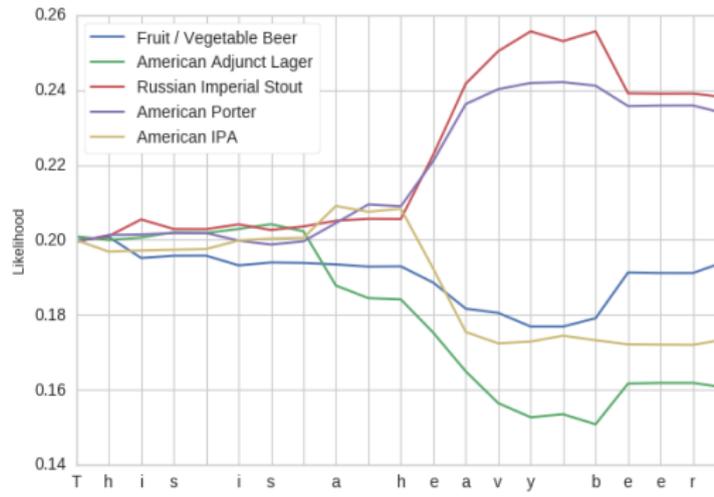


- ▶ Concatenation: input + contexte
- ▶ Inférence = test (combinatoire) des contextes + max de vraisemblance

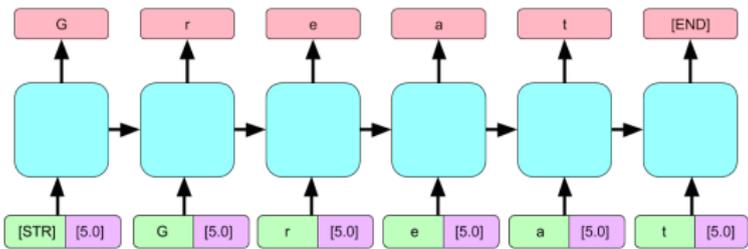
Démo: <http://deepx.ucsd.edu/>



Lipton, Vikram, McAuley, arXiv, 2016
Gen. Concatenative Nets Jointly Learn to Write and Classify Reviews



RNN & modélisation du contexte



- ▶ Concatenation: input + contexte
- ▶ Inférence = test (combinatoire) des contextes + max de vraisemblance

Perspective:

- ▶ Authorship
- ▶ ⇒ Réduire la combinatoire...
- ▶ ... Ou traiter des contextes continus

Démo: <http://deepx.ucsd.edu/>

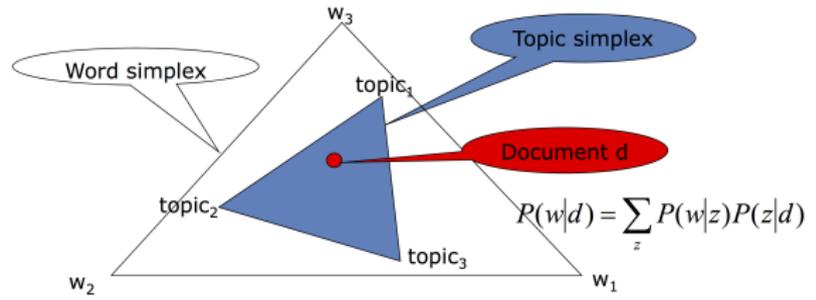
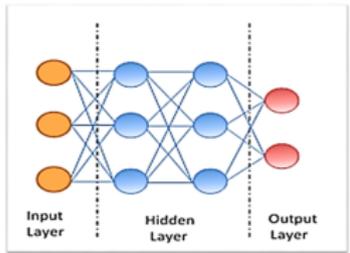


Lipton, Vikram, McAuley, arXiv, 2016
Gen. Concatenative Nets Jointly Learn to Write and Classify Reviews

Apprentissage de représentations:

1 objet brut \Rightarrow 1 représentation vectorielle dont les variables latentes correspondent à des **concepts de haut niveau**

Vision 1990-2000 : extraction de caractéristiques / thématiques



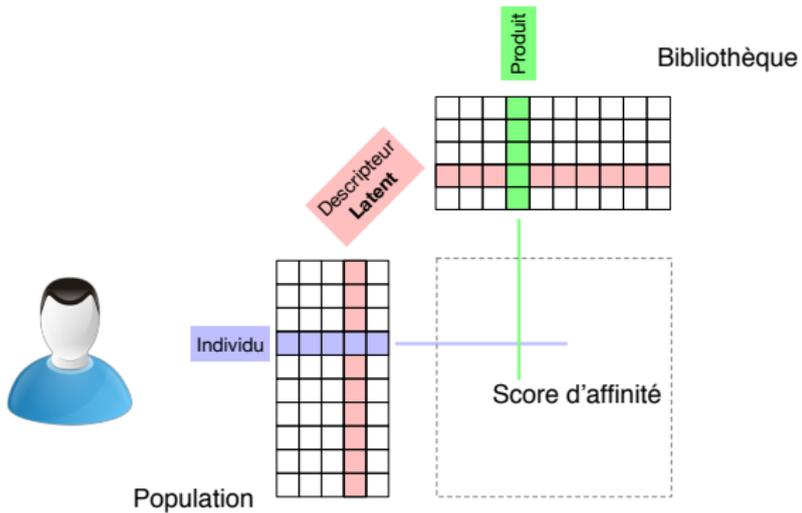
Apprentissage de représentations:

1 objet brut \Rightarrow 1 représentation vectorielle dont les variables latentes correspondent à des **concepts de haut niveau**

Vision 2000-2010 : apprentissage de profils (utilisateurs)

Factorisation matricielle / compression d'information

- Compression d'une matrice de notes
- ▶ Completion : *qui aime quoi ?*
 - ▶ Explication : *qui possède quel facteur ?*
 - ▶ Communautés





Les enjeux de l'apprentissage de représentations

Apprendre une représentation numérique pour les mots, les phrases, les documents...

- ▶ Encoder une signification / une connaissance sur un point de l'espace
 - Nouvelle manière de penser l'extraction d'information
 - Nouvelle approche de la traduction automatique
 - Supervision plus légère / plus de ressources disponibles
 - Traductions plus intelligibles

- ⇒ Les tâches de POS-tagging / SRL ont-elles encore une utilité?

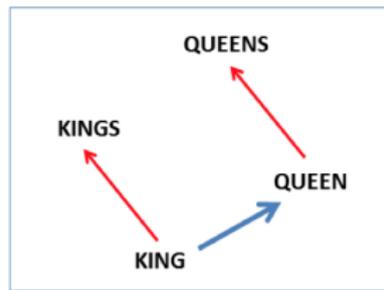
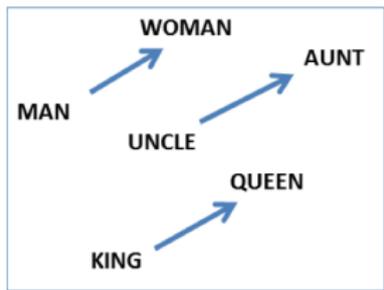
- ▶ Systèmes end-to-end
 - chatbot



Nouvelles propriétés des espaces de mots : Word2Vec

a est à b ce que c est à ??? $\Leftrightarrow \mathbf{z}_b - \mathbf{z}_a + \mathbf{z}_c$

Propriété syntaxique (1):



$$\mathbf{z}_{woman} - \mathbf{z}_{man} \approx \mathbf{z}_{queen} - \mathbf{z}_{king}$$

$$\mathbf{z}_{kings} - \mathbf{z}_{king} \approx \mathbf{z}_{queens} - \mathbf{z}_{kings}$$

Requête:

$$\mathbf{z}_{woman} - \mathbf{z}_{man} + \mathbf{z}_{king} = \mathbf{z}_{req}$$

Plus proche voisin:

$$\arg \min_i \|\mathbf{z}_{req} - \mathbf{z}_i\| = queen$$

⇒ Mise en place d'un test quantitatif de référence.



Nouvelles propriétés des espaces de mots : Word2Vec

a est à b ce que c est à ??? $\Leftrightarrow \mathbf{z}_b - \mathbf{z}_a + \mathbf{z}_c$

Propriété syntaxique (2):

Requête:

$$\mathbf{z}_{easy} - \mathbf{z}_{easiest} + \mathbf{z}_{luckiest} = \mathbf{z}_{req}$$

Plus proche voisin:

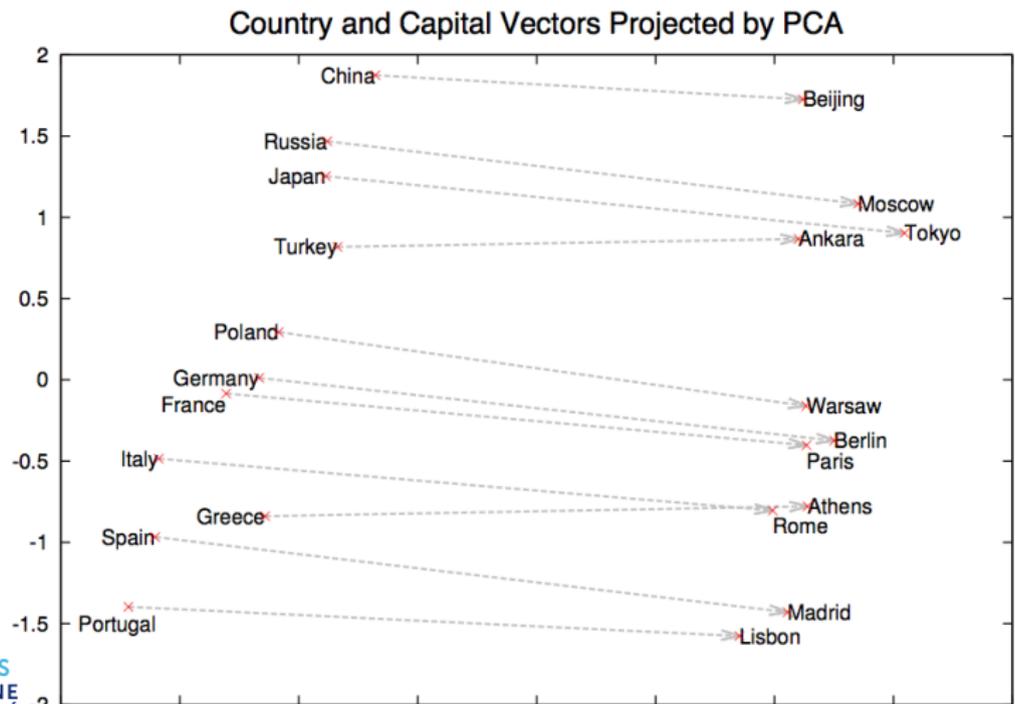
$$\arg \min_i \|\mathbf{z}_{req} - \mathbf{z}_i\| = \text{lucky}$$



Nouvelles propriétés des espaces de mots : Word2Vec

a est à b ce que c est à ??? $\Leftrightarrow \mathbf{z}_b - \mathbf{z}_a + \mathbf{z}_c$

Propriété sémantique (1)



LI^B Nouvelles propriétés des espaces de mots : Word2Vec

a est à b ce que c est à ??? $\Leftrightarrow \mathbf{z}_b - \mathbf{z}_a + \mathbf{z}_c$

Propriété sémantique (2)

Czech + currency	Vietnam + capital	German + airlines	Russian + river	French + actress
koruna	Hanoi	airline Lufthansa	Moscow	Juliette Binoche
Check crown	Ho Chi Minh City	carrier Lufthansa	Volga River	Vanessa Paradis
Polish zolty	Viet Nam	flag carrier Lufthansa	upriver	Charlotte Gainsbourg
CTK	Vietnamese	Lufthansa	Russia	Cecile De

Table 5: Vector compositionality using element-wise addition. Four closest tokens to the sum of two vectors are shown, using the best Skip-gram model.

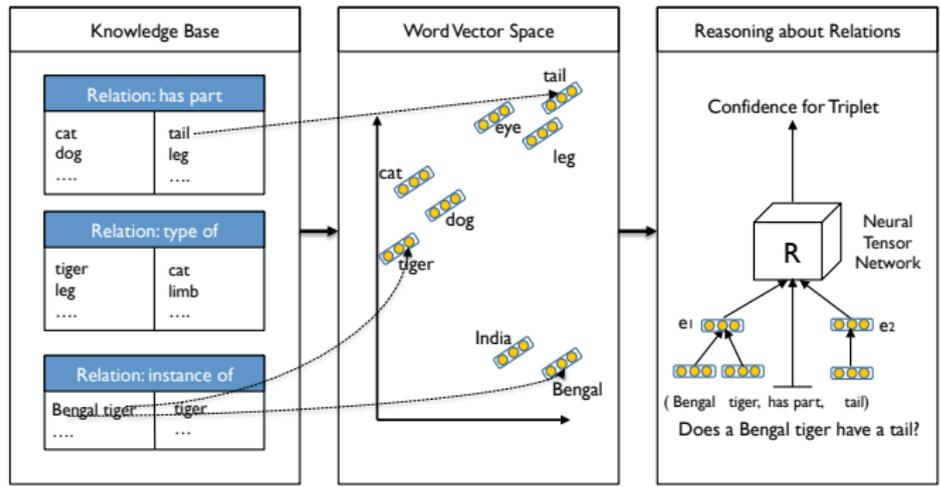


Raisonnement automatique

< 2008

- ▶ Systèmes logiques (prolog)
- ▶ Application de règles
- ▶ Analyse morpho-syntaxique des phrases
- ▶ Classifieurs locaux & caractéristiques manuelles

> 2008

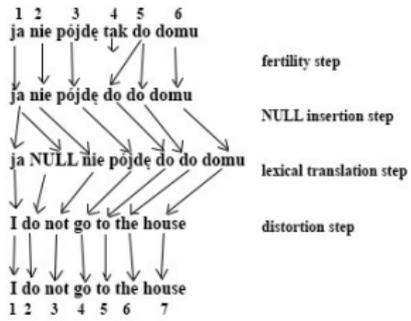




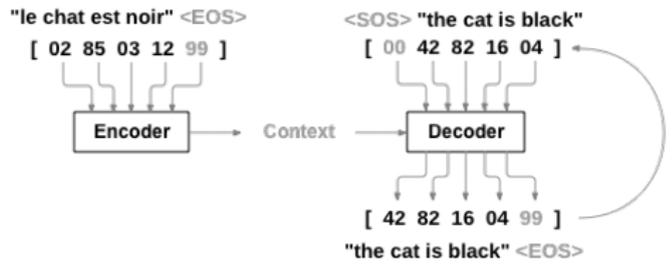
Traduction automatique



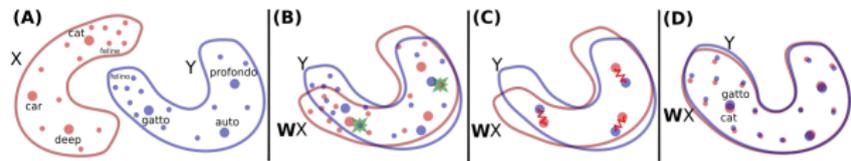
Ancien paradigme



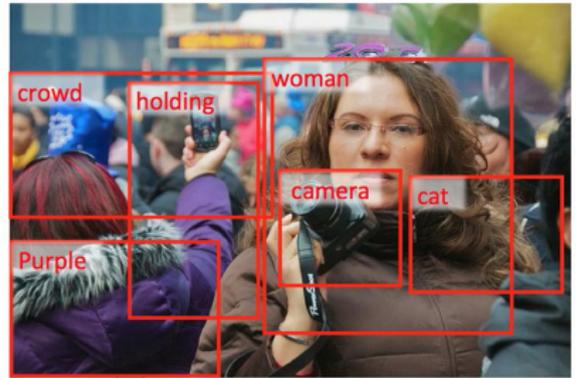
Paradigme RNN



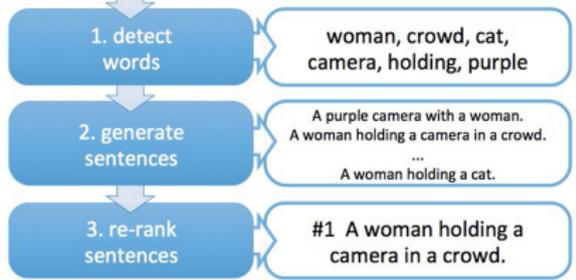
De moins en moins de ressources requises:



LPB Vision & texte : description image



Couplage de la détection d'objet et de la génération de texte



Microsoft

Accès à l'information : des moteurs de recherche aux chatbots

Système d'accès à l'information personnalisé avec 2 particularités:

Dynamique Le profil de l'utilisateur évolue au fil du dialogue

Dialog State Tracking

End to end Requête = langage naturel, Réponse = langage naturel

