Étiquetage syntaxique

- En plus de l'identité des mots, il peut être utile de connaître l'**étiquette syntaxique** de chacun de ces mots
  - « une visite à la ferme » → « ferme » est un nom
  - « Jean ferme la porte » → « ferme » est un verbe

- Connaître la catégorie grammaticale d'un mot peut faciliter une autre tâche
  - ex.: traduction automatique
    - si « ferme » est un nom → « *farm* »
    - si « ferme » est un verbe → « *close* »

Étiquetage syntaxique

- On suppose qu'on a accès à T **corpus étiquetés** D_t (pour simplifier: un document = une phrase)

w_t	e_t
Jean	Nom
ferme	Verbe
la	Article
porte	Nom
.	.

$$D_t = [\ (w^t_1, e^t_1), \ldots, (w^t_d, e^t_d)\ ]$$

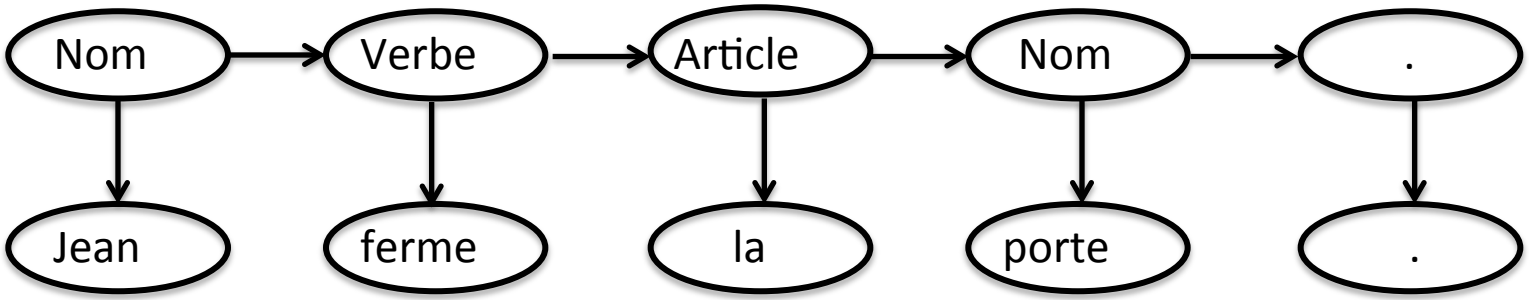
$$\text{mots}(D_t) = [\ w^t_1, \ldots, w^t_d\ ]$$

$$\text{étiqettes}(D_t) = [\ e^t_1, \ldots, e^t_d\ ]$$

- On pourrait prendre une approche similaire à la classification de documents
  - définir un réseau bayésien sur les mots et les étiquettes
  - apprendre le réseau sur notre corpus étiqueté
  - pour faire des prédictions, faire de l'inférence dans le réseau bayésien

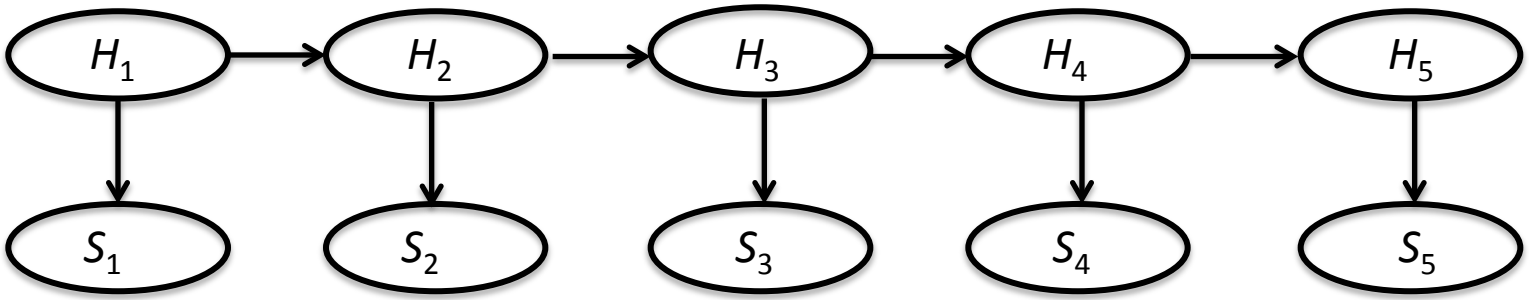
Étiquetage syntaxique par HMM

- On va utiliser un modèle de Markov caché (HMM)



Étiquetage syntaxique par HMM

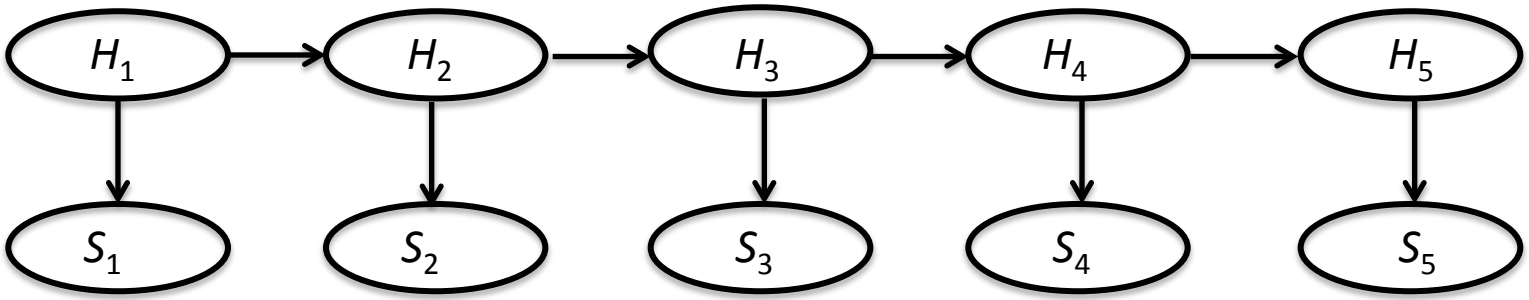
- On va utiliser un modèle de Markov caché (HMM)



  - De notre corpus d'entraînement, on peut extraire des statistiques
    - » sur la première étiquette d'une phrase ($P(H_1)$)
    - » sur la relation entre un mot et sa classe syntaxique ($P(S_k|H_k)$)
      - – ex.: « ferme » peut être un nom, un verbe, mais pas un article
    - » sur la relation entre les étiquettes syntaxiques adjacentes ($P(H_{k+1}|H_k)$)
      - – ex.: on ne peut avoir deux articles qui se suivent

Étiquetage syntaxique par HMM

- On apprend le HMM à partir de ces statistiques (fréquences)

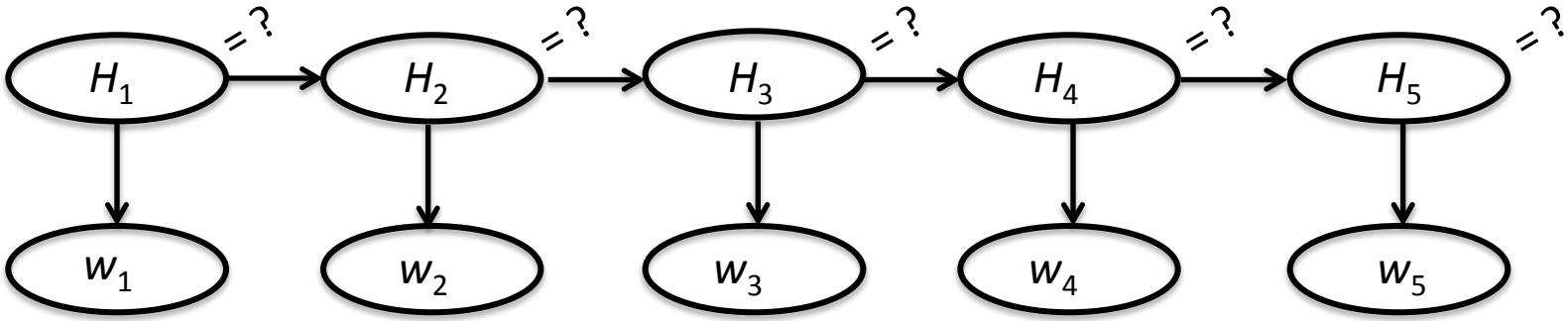


$$P(H_{k+1} = a \mid H_k = b) = \frac{\sum_t \text{freq}(\ (b,a), \text{étiquettes}(D_t)\ )}{\sum_t \text{freq}(\ (b,*), \text{étiquettes}(D_t)\ )} \qquad P(S_k = w \mid H_k = b) = \frac{\sum_t \text{freq}(\ (w,b), D_t\ )}{\sum_t \text{freq}(\ (*,b), D_t\ )}$$

$$P(H_1 = a) = \frac{\sum_t \text{freq}(\ e^t{}_1 = a, D_t\ )}{T}$$

Étiquetage syntaxique par HMM

- Pour étiqueter une nouvelle phrase [w_1, ..., w_d]



- On calcule l'explication la plus plausible $h^*_{1:d}$
  - c.-à-d. $h^*_{1:d}$ qui maximise $P(H_{1:d} = h^*_{1:d}, S_{1:d} = [\, w_1, ..., w_d \,])$
  - on utilise le programme dynamique de α*
    (capsules sur réseaux bayésiens dynamiques)