Extraction d'information

- Nous avons vu comment catégoriser des documents
- Nous avons vu comment les étiqueter automatiquement

- Une fois un document trouvé, comment y extraire l'information désirée automatiquement?

- Exemple: extraire l'information d'une annonce de séminaire
  - le nom du présentateur
  - la date de la présentation

« *There will be a seminar by Dr. Andrew McCallum on Friday* »

Présentateur: *Dr. Andrew McCallum*
Date: *Friday* (vendredi)

Extraction d'information

- On peut aussi **formuler comme un problème d'étiquetage** de mots!
  - il y a **4 étiquettes**
    - **PRE** : préambule de l'information cherchée
    - **TARGET** : l'information à extraire
    - **POST** : fin de l'information
    - **-** : autres mots

Text:	There	will	be	a	seminar	by	Dr.	Andrew	McCallum	on	Friday
Speaker:	-	-	-	-	PRE	PRE	TARGET	TARGET	TARGET	POST	-
Date:	-	-	-	-	-	-	-	-	-	PRE	TARGET

- On entraînerait un HMM par information recherchée
  - HMM « présentateur »
  - HMM « date »

Extraction d'information

- On peut aussi **formuler comme un problème d'étiquetage** de mots!
  - il y a **4 étiquettes**
    - » **PRE** : préambule de l'information cherchée
    - » **TARGET** : l'information à extraire
    - » **POST** : fin de l'information
    - » **-** : autres mots

Text:	There	will	be	a	seminar	by	Dr.	Andrew	McCallum	on	Friday
Speaker:	-	-	-	-	PRE	PRE	TARGET	TARGET	TARGET	POST	-
Date:	-	-	-	-	-	-	-	-	-	PRE	TARGET

- Pour le HMM « présentateur », le corpus d'entraînement contiendrait

[(« There », -), (« will », -), (« be », -), (« a », -), (« seminar », PRE), (« by », PRE),
(« Dr. », TARGET), (« Andrew », TARGET), (« McCallum », TARGET), (« on », POST), (« Friday », -)]

Extraction d'information

- On peut aussi **formuler comme un problème d'étiquetage** de mots!
  - il y a **4 étiquettes**
    - **PRE** : préambule de l'information cherchée
    - **TARGET** : l'information à extraire
    - **POST** : fin de l'information
    - **-** : autres mots

Text:	There	will	be	a	seminar	by	Dr.	Andrew	McCallum	on	Friday
Speaker:	-	-	-	-	PRE	PRE	TARGET	TARGET	TARGET	POST	-
Date:	-	-	-	-	-	-	-	-	-	PRE	TARGET

- Pour le HMM « date », le corpus d'entraînement contiendrait

[(« There », -), (« will », -), (« be », -), (« a », -), (« seminar », -), (« by », -),
(« Dr. », -), (« Andrew », -), (« McCallum », -), (« on », PRE), (« Friday », TARGET)]

Extraction d'information

- On peut aussi **formuler comme un problème d'étiquetage** de mots!
  - il y a **4 étiquettes**
    - » **PRE** : préambule de l'information cherchée
    - » **TARGET** : l'information à extraire
    - » **POST** : fin de l'information
    - » **-** : autres mots

Text:	There	will	be	a	seminar	by	Dr.	Andrew	McCallum	on	Friday
Speaker:	-	-	-	-	PRE	PRE	TARGET	TARGET	TARGET	POST	-
Date:	-	-	-	-	-	-	-	-	-	PRE	TARGET

- Étant donnée une nouvelle phrase
  - l'explication la plus plausible calculée à partir du HMM « présentateur » permettrait d'isoler l'information sur le présentateur
  - l'explication la plus plausible calculée à partir du HMM « date » permettrait d'isoler l'information sur la date de présentation