Exercices : traduction et information mutuelle

Caio Corro

1 Comptage et fréquences

Soit le corpus de phrases parallèles suivant :

	Français	**Anglais**
1	La cuisine est en haut.	The kitchen is upstairs
2	Je vais à la cuisine.	I am going to the kitchen.
3	La salle de bain est là bas.	The bathroom is there
4	Je ne cuisine pas très bien.	I don't cook very well,
5	Le chien est dans le jardin	The dog is in the garden.
6	Où est la cuisine ?	Where is the kitchen
7	Le gateau est dans la cuisine.	The cake is in the kitchen.
8	La porte est ouverte.	The door is open.
9	La chambre est là bas.	The bedroom is there.
10	Le restaurant s'appelle "The vegetarian kitchen"	The restaurant is called "The vegetarian kitchen"

On utilisera les notations suivantes :

- $\#(X)$: nombre de phrase en Français contenant le mot "cuisine" ;
- $\#(Y)$: nombre de phrase en Anglais contenant le mot "kitchen" ;
- $\#(X\&Y)$: nombre de phrases parallèles Français/Anglais où "cuisine" apparaît dans la phrase en Français et "kitchen" apparaît dans la phrase en Anglais ;
- n : nombre total de phrases en Français (ou en Anglais).

On va également utiliser X et Y comme des variables aléatoires avec :

- $P(X = 1)$: probabilité **d'observer** le mot "cuisine" lorsque l'on tire une phrase en Français au hasard dans le corpus ;
- $P(X = 0)$: probabilité **de ne pas observer** le mot "cuisine" lorsque l'on tire une phrase en Français au hasard dans le corpus ;
- $P(Y = 1)$: probabilité **d'observer** le mot "kitchen" lorsque l'on tire une phrase en Anglais au hasard dans le corpus ;
- $P(Y = 0)$: probabilité **de ne pas observer** le mot "kitchen" lorsque l'on tire une phrase en Anglais au hasard dans le corpus ;
- $P(X = 1, Y = 1)$: probabilité **d'observer** "cuisine" dans la phrase en Français et "kitchen" dans la phrase en Anglais quand on tire un couple de phrases alignées au hasard dans le corpus ;
- $P(X = 1, Y = 0)$: probabilité **d'observer** "cuisine" dans la phrase en Français et **de ne pas observer** "kitchen" dans la phrase en Anglais quand on tire un couple de phrases alignées au hasard dans le corpus ;
- etc etc.

Questions :

1. Calculer les valeurs de $\#(X)$, $\#(Y)$ et $\#(X\&Y)$.
2. Calculer toutes les probabilités suivantes en utilisant seulement $\#(X)$, $\#(Y)$, $\#(X\&Y)$ et n :
   - $P(X = 1)$
   - $P(X = 0)$
   - $P(Y = 1)$
   - $P(Y = 0)$
   - $P(X = 0, Y = 0)$
   - $P(X = 1, Y = 0)$
   - $P(X = 0, Y = 1)$
   - $P(X = 1, Y = 1)$

2 Information mutuelle

Soit deux variables aléatoires X et Y et deux distribution P et Q définies de la façon suivante :

- Distribution P :
  - $P(X = 0, Y = 0) = 0.4$
  - $P(X = 1, Y = 0) = 0.1$
  - $P(X = 0, Y = 1) = 0.1$
  - $P(X = 1, Y = 1) = 0.4$
- Distribution Q :
  - $P(X = 0, Y = 0) = 0.1$
  - $P(X = 1, Y = 0) = 0.4$
  - $P(X = 0, Y = 1) = 0.4$
  - $P(X = 1, Y = 1) = 0.1$

Questions :

1. Calculer les distributions marginales de X et Y pour les distributions jointes P et Q.
2. Calculer $P(Y = 1|X = 1)$
3. Calculer $P(Y = 0|X = 1)$
4. Calculer $Q(Y = 1|X = 1)$
5. Calculer $Q(Y = 0|X = 1)$
6. Calculer l'information mutuelle de X et Y pour les distributions P et Q
7. Comment interpréter les résultats précédents ? Que cela implique-t-il pour notre méthode d'extraction automatique d'un dictionnaire de traduction ?