

TD3 - Expériences et données

Carlos Ramisch et Manon Scholivet - Recherche Zen

L'objectif de cette activité est de vous familiariser par la pratique avec certains outils de la préparation d'expériences, notamment la création et évaluation de jeux de données, ainsi que la conception du protocole et des mesures d'évaluation. Les activités ci-dessous sont relativement indépendantes les unes des autres (sauf les deux premières) et seront effectuées tout au long du cours.

1 Annotation de données (20 min)

L'annotation de données pose bien de nombreux problèmes qu'il est parfois difficile d'appréhender sans les avoir expérimentés. Nous vous proposons de vous familiariser avec une tâche d'annotation, celle des expressions polylexicales (EPL). Une EPL est un groupe d'unités lexicales qui doivent être considérées comme une unique unité lexicale. Par exemple : "C'est la *douche froide*."

Rendez-vous sur le lien : https://parsemefr.lis-lab.fr/compounds_fr/.

Et suivez les instructions !

2 Accord inter-annotateur.ice (20 min)

Après avoir annoté une vingtaine d'expressions, vous devriez être plus familier avec cette tâche. Mais avez-vous annoté de façon cohérente avec le reste du groupe ?

Pour répondre à cette question, nous allons calculer le score d'accord inter-annotateur.ice.

Pour cela, vous avez pu voir que le score Kappa se calcule ainsi :

- Proportion d'accord au dessus de l'aléatoire :

$$\kappa = \frac{A_O - A_E}{1 - A_E}$$

- Accord "par chance" : suppose que chaque annotateur.ice a sa propre distribution

$$A_E^\kappa = \frac{1}{i^2} \sum_{k \in K} n_{c_1 k} n_{c_2 k}$$

- i nombre total d'items annotés (ici, 20),
- K nombre de valeurs par item (ici, 6 catégories),
- $n_{c_j k}$ nombre d'items annotés k par l'annotateur.ice c_j

Étant donné les scores d'annotations que nous allons vous fournir, calculer le score kappa du groupe. Pour récupérer la totalité des données annotées, vous pouvez aller sur la page https://parsemefr.lis-lab.fr/compounds_fr/howisitgoing.php. Nous vous conseillons de copier-coller les données sur un tableur puis les transformer en tableau avec les EPL sur les lignes et les annotateur.ice.s sur les colonnes. Nous allons calculer le score κ uniquement pour la réponse à la question finale, *answer-headModifier*.

3 Métriques d'évaluation (30 min)

Après avoir annoté des EPL vient l'étape de leur prédiction dans un nouveau texte! Mais comment vérifier la qualité de la prédiction ?

Il va être nécessaire de créer une mesure d'évaluation. Mais vous allez rapidement être confrontés à un certain nombre de problèmes. En voici quelques uns :

- Les EPL peuvent être discontinues ! Par exemple : "Il *prend* ces problèmes *en compte*."
- Que faire si seulement une partie de l'EPL est prédite ?
- Que faut-il évaluer ? La précision ? Le rappel ? Autre chose ?
- Faut-il accorder plus de valeur à une EPL qui n'a pas été vue dans le corpus d'entraînement ?

Pour vous guider, vous pouvez utiliser un exemple comme celui-ci, qui illustre certains des défis listés ci-dessus :

Gold: make segmentation decisions, in order to split sentences into lexical units

System: make segmentation decisions, in order to split sentences into lexical units

À vous de trouver la (ou les ?) meilleure(s) métrique(s) ! Essayer de formuler la métrique de manière assez précise, voire la formaliser avec des équations si vous y arrivez.

4 Protocole expérimental (30 min)

Vous rappelez vous de votre question de recherche définie lors du premier cours ? C'est le moment de la ressortir, et d'établir un protocole expérimental qui vous permettra d'y répondre!