Réunion 17/05/2023

NER avec GPT3.5

Valentin BONSI

NER avec GPT3.5

Méthode	Nb erreur max	Nb erreur total	Taux d’erreur pour 2200 labels :
One-shot	13	127	5.77%
Few-shot	7	35	1.5%
Few-shot-corrigé	5	15	0.68%

NER avec GPT3.5

Réceptions des nouvelles données corrigées (test et train) :
nouvelles données = nouveaux problèmes

1. **Actes parfois tronqués**
2. **Actes de divorces**
3. **Possibles erreurs restantes**
4. **Formulation pas encore rencontrées**

NER avec GPT3.5 - Actes tronqués

On ne veut pas :

- Un découpage en 5 paragraphes différents de d'habitude
- Renvoyer moins de 5 paragraphes

On veut :

P1 : {}, P2 : {}, P3 : {…texte}, P4 : {texte}, P5 : {texte}
P1 : {texte}, P2 : {texte}, P3 : {texte…}, P4 : {}, P5 : {}

Solution : Ajouter des exemples :

- Un exemple complet
- Un exemple avec que le début
- Un exemple avec que la fin

NER avec GPT3.5 - Actes tronqués

Exemple :
P1 {}
P2 {}
P3 {sa veuve domicilée au 7 rue des Prés d'autre part}
P4 {…**Marius Fernand VIDAL** et **Charlotte Lucie LEPAGE** ont déclaré…}
P5 {[Les témoins et l'adjoint]}

Avec un acte tronqué des infos recherchées dans un paragraphe peuvent ne pas être présentent.

Mais elles peuvent l'être dans un autre. C'est le cas des coordonnées des mariés

-> On ne laisse plus systématiquement le 4eme paragraphe de côté.

SI et seulement si il nous manque un nom ou prénom, alors on les récupère grâce au paragraphe 4

NER avec GPT3.5 - Nouvelle structure exemples

Plûtot que d'envoyer à l'API un message avec notre question et nos exemples, on lui envoie un 'début' de Chat entre lui et nous :

Exemple 1
Labels 1

Exemple 2
Labels 2

Exemple 3
Labels 3

Exemple 4
Labels 4

Question

Réponse de l'API

System : You are …

Exemple 1

Labels 1

Exemple 2

Labels 2

Exemple 3

Labels 3

Question

Réponse de l'API

NER avec GPT3.5 - Erreurs présentes

Il peut rester des erreurs dans le fichier corrigé. Elles peuvent venir :

- D’erreurs passées au travers de la correction
- De formulation complexes ou inhabituelles mettant à mal le script de conversion du texte annoté

-> Peut-on utiliser GPT pour corriger ces erreurs ? En ajoute-t-on plus que ce qu’on retire ?

NER avec GPT3.5 - Données de Test

1	Nom de l'archive	Nb erreur	Nb erreur corrigées	Nb erreur ajoutées
2	archives_AD075EC_01M1940_0016-left.png-0	5	2	3
3	archives_AD075EC_01M1940_0016-left.png-1	1	0	1
4	archives_AD075EC_01M1940_0016-left.png-2	2	2	0
5	archives_AD075EC_01M1940_0016-right.png-2	4	2	2
6	archives_AD075EC_01M1940_0016-right.png-1	9	9	0
7	archives_AD075EC_11M549_0104-left.png-1	8	5	3
8	archives_AD075EC_11M549_0104-left.png-0	12	0	12
9	archives_AD075EC_11M549_0104-left.png-2	3	1	2
10	archives_AD075EC_11M549_0104-right.png-2	10	7	3
11	archives_AD075EC_01M1940_0024-left.png-0	5	5	0
12	archives_AD075EC_01M1940_0024-left.png-1	2	2	0
13	archives_AD075EC_01M1940_0024-left.png-2	3	3	0
14	archives_AD075EC_01M1940_0024-right.png-2	6	5	1
15	archives_AD075EC_01M1930_0007-left.png-0	6	6	0
16	archives_AD075EC_01M1930_0007-left.png-1	1	1	0
17	archives_AD075EC_01M1930_0007-left.png-2	2	1	1
18	archives_AD075EC_01M1930_0007-right.png-0	4	2	2
19	archives_AD075EC_01M1930_0007-right.png-1	8	0	8
20	archives_AD075EC_01M1930_0007-right.png-2	3	2	1
21	archives_AD075EC_01M1940_0018-left.png-0	7	7	0