Campagne d'évaluation Entités Nommées

- **Ce TP noté constituera la plus grande partie de votre évaluation *Contrôle Continu* de cette matière. Vous vous retrouvez dans la situation de participer à une mini campagne d'évaluation portant sur la détection automatique d'entités nommées dans des textes.**
- **Pour ce TP vous aurez absolument besoin des programmes réalisés durant les deux dernières séances de TP portant sur l'extraction de patrons de POS pour représenter des entités nommées (TP4) et la classification en labels d'entités nommées avec *icsiboost* (TP5).**
- **Vous disposez de nouvelles ressources (corpus d'apprentissage, dictionnaire) qui vont vous permettre d'améliorer les modèles de classification que vous avez pu réaliser au TP4.**
- **Vous disposez également d'un corpus nouveau, composé uniquement des mots et des POS, sans les labels en entités nommées, que vous devrez annoter avec votre meilleur modèle, et le rendre impérativement à la fin du TP de l'après-midi.**
- **Les scores en *macro-F1* et *micro-F1* que vous obtiendrez à l'issue de ce TP rentreront en compte dans votre évaluation.**
- **Vous devrez rendre également un rapport décrivant vos expériences. Une première version est à rendre à la fin de la journée, une deuxième version pourra être rendu jusqu'au 14 février, jour de l'examen, et pour cette version vous aurez les références des fichiers de test que vous devrez traiter.**
- **Voici une description du processus que vous allez devoir mettre en place :**

Tout d'abord vous récupèrerez un corpus au même format que *corpus_en_200k.train.txt* (mais plus grand, appelé : *corpus_en_580K.train* dont vous allez extraire un corpus d'apprentissage et l'un de développement) qui contiendra des annotations en POS et entités nommées comme dans l'exemple suivant nommé *corpus_en_exemple.txt*:

```
0    Clinton         np          B-person
1    a               v           O
2    regretté        v           O
3    les             det         O
4    pertes          nc          O
5    civiles         adj         O
6    ,               ponctw      O
7    que             csu         O
8    nous            cln         O
9    ne              advneg      O
10   voulons         v           O
11   pas             advneg      O
12   .               ponctw      O

0    les             det         O
1    cibles          nc          O
2    choisies        v           O
3    cette           det         O
4    nuit            nc          O
5    n'              advneg      O
6    ont             v           O
7    pas             advneg      O
8    été             v           O
9    identifiées     v           O
10   par             prep        O
11   le              det         O
12   Pentagone       np          B-org
13   ,               ponctw      O
14   mais            coo         O
15   le              det         O
16   conseiller      nc          O
17   du              prep        O
18   président       nc          O
19   Sandy           np          B-person
```

```
20   Berger          np            I-person
21   rappelle        v             O
22   l'              det           O
23   objectif        nc            O
24   ,               ponctw        O
25   casser          v             O
26   l'              det           O
27   appareil        nc            O
28   de              prep          O
29   Production      np            O
30   d'              prep          O
31   armes           nc            O
32   de              prep          O
33   destructions    nc            O
34   massives        adj           O
35   de              prep          O
36   Saddam          np            B-person
37   Hussein         np            I-person
38   en              prep          O
39   Irak            np            B-geoloc
40   REUTER          np            O
```

-
- **Ensuite, grâce aux programmes du TP4, vous déterminerez quels sont les patrons de POS les plus fréquents qui peuvent correspondre aux entités nommées de ce corpus. Sur le corpus précédent, *corpus_en_exemple.txt*, il y a 2 patrons possibles : *np* et *np np*.**

En sélectionnant un ensemble de patrons parmi tout ceux possibles, vous utiliserez le programme de l'exercice 3 du TP4 pour marquer avec le symbole *hyp* les séquences de mots candidates pour représenter des entités nommées. Par exemple, avec les 2 patrons *np* et *np np*, nous obtiendrons sur le corpus précédent :

```
0    Clinton         np            B-person     hyp
1    a               v             O            O
2    regretté        v             O            O
3    les             det           O            O
4    pertes          nc            O            O
5    civiles         adj           O            O
6    ,               ponctw        O            O
7    que             csu           O            O
8    nous            cln           O            O
9    ne              advneg        O            O
10   voulons         v             O            O
11   pas             advneg        O            O
12   .               ponctw        O            O

0    les             det           O            O
1    cibles          nc            O            O
2    choisies        v             O            O
3    cette           det           O            O
4    nuit            nc            O            O
5    n'              advneg        O            O
6    ont             v             O            O
7    pas             advneg        O            O
8    été             v             O            O
9    identifiées     v             O            O
10   par             prep          O            O
11   le              det           O            O
12   Pentagone       np            B-org        hyp
13   ,               ponctw        O            O
14   mais            coo           O            O
15   le              det           O            O
16   conseiller      nc            O            O
```

```
17   du             prep          O           O
18   président      nc            O           O
19   Sandy          np            B-person    hyp
20   Berger         np            I-person    hyp
21   rappelle       v             O           O
22   l'             det           O           O
23   objectif       nc            O           O
24   ,              ponctw        O           O
25   casser         v             O           O
26   l'             det           O           O
27   appareil       nc            O           O
28   de             prep          O           O
29   Production     np            O           hyp
30   d'             prep          O           O
31   armes          nc            O           O
32   de             prep          O           O
33   destructions   nc            O           O
34   massives       adj           O           O
35   de             prep          O           O
36   Saddam         np            B-person    hyp
37   Hussein        np            I-person    hyp
38   en             prep          O           O
39   Irak           np            B-geoloc    hyp
40   REUTER         np            O           hyp
```

- **L'étape suivante consiste à produire des données (apprentissage ou test) pour *icsiboost*, en choisissant un ensemble de traits pour décrire les entités nommées, comme dans l'exercice 4 du TP5. Bien sûr uniquement les mots marqués par le label *hyp* deviendront des données pour *icsiboost*. La première colonne devra obligatoirement contenir l'indice du mot dans le fichier d'origine. Cette indice nous permettra de placer les labels d'entités nommées prédits au bon endroit. Ensuite vous pouvez choisir l'ensemble de traits de votre choix en utilisant les mots, les étiquettes morphosyntaxiques (les POS), des caractéristiques sur les mots, des ressources externes, etc. Par exemple, à partir du corpus précédent *corpus_en_exemple.txt*, on peut générer le corpus *corpus_en_exemple.icsiboost* suivant où chaque exemple est représenté par 4 champs : l'indice du mot dans le fichier d'origine, un champs texte représentant le contexte gauche du mot à étiqueter (limité aux 2 mots précédents), puis le mot à étiqueter, puis le contexte gauche (limité aux 2 mots suivants).**

```
0 , XX XX , Clinton, a regretté , B-person .
26 , par le , Pentagone, !VIRGULE mais , B-org .
33 , du président , Sandy, Berger rappelle , B-person .
34 , président Sandy, Berger, rappelle l' , I-person .
43 , appareil de , Production , d' armes , O .
50 , massives de , Saddam, Hussein !POINT , B-person .
51 , de Saddam, Hussein, !POINT XX , I-person .
53 , Hussein en , Irak , REUTER XX , B-geoloc .
54 , en Irak , REUTER , XX XX , O .
```

- **Une fois l'apprentissage d'un modèle fait avec *icsiboost*, vous pourrez déjà mesurer les performances du classifieurs sur les candidats entités nommées sélectionnés grâce aux patrons. Plus vous uttliserez de patrons, plus vous aurez d'exemples d'apprentissage, mais plus vous aurez aussi d'exemples négatifs (label *O*). A vous de trouvez le bon compromis !!**
- **L'évaluation lors de l'apprentissage avec *icsiboost*, même sur des corpus *dev* et *test*, ne suffit pas. En effet, comme vous êtes limité aux seuls mots sélectionnés par vos patrons de POS, vous n'avez pas une évaluation sur l'ensemble des mots et des entités du corpus. Pour faire cette évaluation globale à partir de la classification avec *icsiboost*, nous vous proposons les commandes suivantes (à compiler en C pour les programmes en C ou à executer en python) et à télécharger à la fin de cette section :**

***tagg_corpus_icsiboost -names <file>.names -corpus <file>* : cette commande prend en entrée la sortie**

d'une étape de classification avec *icsiboost -C*, deux arguments correspondant au fichier *names* et au fichier *corpus* qui a été traité par *icsiboost*, et il produit pour chaque exemple à classer, un triplet composé du numéro de ligne de l'exemple dans le fichier original, du label choisi en prenant le score de classification le plus grand sur tous les labels, et enfin du score lui-même. Par exemple, sur le fichier d'exemple précédent *corpus_en_exemple.icsiboost*, en considérant que le fichier *names* utilisé lors de l'apprentissage s'appelle *baseline_en.names*, après execution de la commande :

```
cat corpus_en_exemple.icsiboost | icsiboost -S baseline_en -C | \
       ./tagg_corpus_icsiboost -names baseline_en.names -corpus corpus_en_exemple.icsiboost
```

nous obtiendrons le fichier suivant :

```
0   B-person    0.000345
26  B-org      -0.000133
33  B-person    0.001153
34  B-org       0.000094
43  B-org       0.000202
50  B-person    0.001031
51  I-person    0.001085
53  B-geoloc    0.002010
54  I-geoloc   -0.000086
```

■

***recopie_label_entite_nommes.py file* : cette commande prend en entrée la sortie de *tagg_corpus_icsiboost* ainsi qu'un paramètre contenant le fichier original avec les 4 champs (indice, mot, POS, entités nommées), il recopie ensuite les labels prédits par *icsiboost* aux bonnes lignes dans le fichier, dans une nouvelle 5e colonne. Par exemple, en executant la commande :**

```
 cat corpus_en_exemple.icsiboost | icsiboost -S baseline_en -C | \
          ./tagg_corpus_icsiboost -names baseline_en.names -corpus
corpus_en_exemple.icsiboost |\
          python3 ./recopie_label_entite_nommes corpus_en_exemple.txt
```

nous obtiendrons le fichier suivant :

```
0    Clinton      np           B-person    B-person
1    a            v            O           O
2    regretté     v            O           O
3    les          det          O           O
4    pertes       nc           O           O
5    civiles      adj          O           O
6    ,            ponctw       O           O
7    que          csu          O           O
8    nous         cln          O           O
9    ne           advneg       O           O
10   voulons      v            O           O
11   pas          advneg       O           O
12   .            ponctw       O           O

0    les          det          O           O
1    cibles       nc           O           O
2    choisies     v            O           O
3    cette        det          O           O
4    nuit         nc           O           O
5    n'           advneg       O           O
6    ont          v            O           O
7    pas          advneg       O           O
8    été          v            O           O
9    identifiées  v            O           O
10   par          prep         O           O
11   le           det          O           O
12   Pentagone    np           B-org       B-org
13   ,            ponctw       O           O
14   mais         coo          O           O
```

```
15   le           det          O           O
16   conseiller   nc           O           O
17   du           prep         O           O
18   président    nc           O           O
19   Sandy        np           B-person    B-person
20   Berger       np           I-person    B-org
21   rappelle     v            O           O
22   l'           det          O           O
23   objectif     nc           O           O
24   ,            ponctw       O           O
25   casser       v            O           O
26   l'           det          O           O
27   appareil     nc           O           O
28   de           prep         O           O
29   Production   np           O           B-org
30   d'           prep         O           O
31   armes        nc           O           O
32   de           prep         O           O
33   destructions nc           O           O
34   massives     adj          O           O
35   de           prep         O           O
36   Saddam       np           B-person    B-person
37   Hussein      np           I-person    I-person
38   en           prep         O           O
39   Irak         np           B-geoloc    B-geoloc
40   REUTER       np           O           I-geoloc
```

- **Dans ce fichier là, le label de référence en entité nommé est dans la colonne 4, et l'hypothèse prédite dans la colonne 5.**

evalue_entite_nomme **: cette commande prend en entrée la sorte de *recopie_label_entite_nommes* et produit une évaluation complète en comparant les labels de référence (colonne 4) et ceux prédits (colonne 5). Par exemple, sur l'exemple précédent, nous obtiendrons :**

```
Evaluation stricte (labels et frontieres doivent etre corrects) :
  - macro-F1 : Precision: 33.33 - Rappel: 55.56 - F1: 41.67 - nbref=5 nbhyp=7 nbok=3
  - micro-F1 : Precision: 42.86 - Rappel: 60.00 - F1: 50.00 - nbref=5 nbhyp=7 nbok=3
  Details par label :
    -     geoloc : Precision: 00.00 - Rappel: 00.00 - F1: 00.00 - nbref=1 nbhyp=1 nbok=0
    -        org : Precision: 33.33 - Rappel: 100.00 - F1: 50.00 - nbref=1 nbhyp=3 nbok=1
    -     person : Precision: 66.67 - Rappel: 66.67 - F1: 66.67 - nbref=3 nbhyp=3 nbok=2

Evaluation detection (uniquement le token de debut et le type de l'entite doivent etre
corrects) :
  - macro-F1 : Precision: 77.78 - Rappel: 100.00 - F1: 87.50 - nbref=5 nbhyp=7 nbok=5
  - micro-F1 : Precision: 71.43 - Rappel: 100.00 - F1: 83.33 - nbref=5 nbhyp=7 nbok=5
  Details par label :
    -     geoloc : Precision: 100.00 - Rappel: 100.00 - F1: 100.00 - nbref=1 nbhyp=1 nbok=1
    -        org : Precision: 33.33 - Rappel: 100.00 - F1: 50.00 - nbref=1 nbhyp=3 nbok=1
    -     person : Precision: 100.00 - Rappel: 100.00 - F1: 100.00 - nbref=3 nbhyp=3 nbok=3
```

Vous pouvez enchaîner toutes ces commandes pour évaluer directement un modèle :

```
cat corpus_en_exemple.icsiboost | icsiboost -S baseline_en -C | \
       ./tagg_corpus_icsiboost -names baseline_en.names -corpus corpus_en_exemple.icsiboost
| \
       python3 ./recopie_label_entite_nommes corpus_en_exemple.txt | ./evalue_entite_nomme
```

- **C'est cette évaluation que vous utiliserez pour mettre au point vos modèles. Vous pourrez jouer à la fois sur les patrons de POS pour avoir plus ou moins de candidats entités nommées, avec plus ou moins d'ambigüité, et aussi sur les traits que vous utiliserez pour le classifieur *icsiboost*.**

- **Travail à rendre - partie 1**
- **Vous allez rendre un rapport détaillant vos expériences pour obtenir le meilleur modèle possible sur votre corpus de développement issus de *corpus_en_580K.train***
    - **Pour chaque expérience vous pouvez faire varier un ou plusieurs de ces éléments : le choix des patrons de POS utilisés pour sélectionner les séquences de mots à étiqueter ; les traits décrivant les données dans les fichiers Icsiboost ; les hyperparamètres d'Icsiboost lors de l'apprentissage.**
    - **Pour le choix des traits vous devrez proposer des modélisations plus riches que celles utilisées dans le TP précédent. Vous pourrez utiliser des champs texte plutôt que des mots isolés pour représenter un contexte plus long ; des caractéristiques sur les mots telles que la présence d'une majuscule en début de phrase ; des ressources externes comme les dictionnaires de noms propres ; des mesures d'ambiguités calculées sur le corpus d'apprentissage, etc.**
    - **Vous sélectionnerez un ensembre d'expériences (entre 4 et 6) parmi toutes celles que vous aurez réalisé, et vous donnerez dans votre rapport les courbes d'apprentissage ainsi que les évaluation en entités nommées sur chaque sortie. Vous commenterez ces résultats dans votre rapport.**
- **Travail à rendre - partie 2**
- **Vous allez maintenant récupérer le fichier d'évaluation qui correspond à votre groupe. Ce fichier s'appelle *corpus_eval_groupe8.eval* si vous êtes dans le groupe 8. C'est un fichier au même format que *corpus_en_580K.train* sauf que la 4e colonne contient uniquement la valeur 'O', et pas les valeurs corrects d'entités nommées que vous devrez prédire.**

Commencez par récupérer votre fichier d'évaluation et traitez le avec votre meilleur modèle obtenu durant la journée avec la même chaîne de traitement que pour le corpus de développement. Par exemple si vous êtes le groupe 8 :

```
 cat corpus_eval_groupe8.icsiboost | icsiboost -S baseline_en -C | \
         ./tagg_corpus_icsiboost -names mon_meilleur_modele.names -corpus
corpus_eval_groupe8.icsiboost |\
```

- ```
  python3 ./recopie_label_entite_nommes corpus_eval_groupe8.eval >
  corpus_eval_groupe8.prediction
  ```
- **Ensuite vous déposerez dans le devoir à rendre aujourd'hui à 18h votre rapport ainsi que votre fichier *corpus_eval_groupe.prediction***
  ```