Faculté des Sciences
Aix*Marseille Université

Année universitaire 2020/2021

Site : ☒ Luminy ☐ St-Charles ☐ St-Jérôme ☐ Cht-Gombert ☐ Aix-Montperrin ☐ Aubagne-SATIS
Sujet de : ☐ 1er semestre ☒ 2ème semestre ☐ Session 2 Durée de l'épreuve : 2h
Examen de : M1 Nom du diplôme : Master Informatique
Code du module : SINBU02L Libellé du module : Introduction au Traitement Automatique des Langues
Calculatrices autorisées : NON Documents autorisés : NON

1 Généralités sur le Traitement Automatique des Langues (7 pts)

Q.1. Pourquoi les langages naturels sont-ils *implicites* et *ambigus* alors que les langages formels telles que les langages informatiques sont *explicites* et *non ambigus* ?

Les langages naturels sont implicites et ambigus pour des soucis d'efficacité : la quantité de connaissances communes entre les humains permet de simplifier le discours et d'éviter de tout préciser. A l'inverse, pour les langages *formels*, seule la forme d'une phrase doit permettre son interprétation. Quelle que soit la machine sur laquelle le programme va s'executer, la suite d'opérations doit produire le même résultat. Ce n'est pas dépendant du contexte ou des connaissances partagées entre les interlocuteurs.

Q.2. On trouve grossièrement deux familles de méthodes pour le TAL : le méthodes à base de *connaissances explicites* et les méthodes *numériques basées sur l'apprentissage automatique*. Citez des avantages et des inconvénients pour ces deux familles de méthodes ? Citez quelques méthodes de chacune de ces familles.

- *connaissances explicites*
  - **Avantages** : explicabilité des décisions prises, contrôle des erreurs ; possibilité de rajouter directement de nouvelles connaissances
  - **Inconvénients** : manque de robustesse (les connaissances doivent couvrir tous les cas d'utilisation possible) ; nécessite une expertise forte
- *méthodes numériques / apprentissage*
  - **Avantages** : robustesse et efficacité pour une tâche donnée, si suffisament de données d'apprentissage sont disponibles
  - **Inconvénients** : difficile d'expliquer les erreurs commises (boite noire) ; nécessite beaucoup de données d'apprentissage

Q.3. Décrivez différentes méthodes de représention d'un mot et d'un texte dans une application de TAL.

mot :
- une entrée dans une base de donnée
- vecteur numérique
- un arbre en partie commune
- ...

phrase :
- sac de token (ensemble)
- séquence de token (liste)
- arbre (suite à une étape d'analyse syntaxique)
- vecteur
- ...

Q.4. Qu'est-ce que la loi de Zipf ?

Voir cours : *Quelques éléments de lexicométrie, analyse statistique de textes*

Q.5. Citez les différents niveaux d'analyse linguistique d'une phrase en TAL

Voir cours : *Cours 3 - Niveaux d'analyse linguistique* - diapos 6 à 15

Q.6. Quels sont les avantages et les inconvénients des architectures *pipeline* pour les chaînes d'analyse linguistique ? (dans un *pipeline* les modules se suivent, la sortie du module n étant l'entrée du module $n+1$)

Voir cours : *Cours 3 - Niveaux d'analyse linguistique* - diapos 24

Q.7. Quelles sont les principales difficultés de la tâche d'étiquetage en *entités nommées* ?

Voir cours : *Entités Nommées* - diapos 4 et 5

2 Evaluation (5 pts)

Soit un système de détection d'entité nommées produisant les résultats suivants :

```
1   Sandy          np      B-person  B-person
2   Berger         np      I-person  I-person
3   rappelle       v       O         B-geoloc
4   l'             det     O         O
5   objectif       nc      O         O
6   ,              ponctw  O         O
7   casser         v       O         O
8   la             det     O         O
9   Production     np      O         B-org
10  d'             prep    O         O
11  armes          nc      O         O
12  de             prep    O         O
13  destructions   nc      O         O
14  massives       adj     O         O
15  de             prep    O         O
16  Saddam         np      B-person  B-person
17  Hussein        np      I-person  I-person
18  en             prep    O         O
19  Irak           np      B-geoloc  B-geoloc
20  .              poncts  O         O
```

L'annotation de référence en entités nommées se trouve en colonne 4 et la prédiction automatique en colonne 5.

Q.8. Ecrivez les formules permettant de calculer le taux d'étiquettes correctes (C) au niveau des mots et la précision (P), le rappel (R) et la F-mesure ($F1$) au niveau des entités nommées.

Soit un texte de N mots, $Nb(mot,correct)$ le nombre de prédiction correcte des étiquettes au niveau des mots, on a : $C = \frac{Nb(mot,correct)}{N}$

Pour chaque entité nommée de type T, on note $Nb(T,reference)$, $Nb(T,prediction)$ et $Nb(T,correct)$, respectivement le nombre d'entités de type T dans l'annotation de référence du texte ; le nombre d'entité de type T prédites par le système automatique que l'on évalue sur le texte ; et enfin le nombre d'entité prédites correctes. Nous avons :

$P_T = \frac{Nb(T,correct)}{Nb(T,prediction)}$ $R_T = \frac{Nb(T,correct)}{Nb(T,reference)}$ $F_T = \frac{2 \times P_T \times R_T)}{P_T + R_T}$

Q.9. Sur l'exemple précédent, donnez les valeurs de C, $P_{\texttt{label}}$, $R_{\texttt{label}}$ et $F1_{\texttt{label}}$ pour les labels `geoloc`, `org`, `person` (évaluation strite des entités nommées).

$C = \frac{18}{20}$ $P_{person} = \frac{2}{2}$ $R_{person} = \frac{2}{2}$ $F_{person} = 1$
$P_{geoloc} = \frac{1}{2}$ $R_{geoloc} = \frac{1}{1}$ $F_{geoloc} = \frac{2}{3}$ $P_{org} = \frac{0}{1}$ $R_{org} = \frac{0}{0}$ $F_{org} = 0$

Q.10. Donnez les valeurs de macro-F1 et micro-F1 obtenues.

Voir cours : *Cours 5 - Evaluation et TAL* - diapo 11

Q.11. Pourquoi la seule métrique C ne suffit-elle pas ?

Dans une tâche telle que les entités nommées, il y a beaucoup trop de 'O' comparé aux étiquettes d'entités. Un système qui prédit toujours 'O' a donc un score C très correct.

Q.12. Soit un système A obtenant un rappel de 70% et une précision de 30%, et un système B obtenant un rappel de 30% et une précision de 70%. Quel est le meilleur système ? Pourquoi ?

Tout dépend du contexte applicatif, on ne peut pas dire qu'un système est meilleur que l'autre dans l'absolu.

3 Questions sur les TP (8 pts)

La figure 1 montre les performances obtenues en classification en étiquettes morphosyntaxiques (POS) durant l'apprentissage de 3 modèles sur 4000 itérations : le *modèle1* ne contient comme traits que le contexte immédiat du mot à étiqueter ; le *modèle2* contient en plus la séquence de lettres de chaque mot ; le *modèle3* contient en plus la liste de étiquettes possibles selon un lexique syntaxique (le `lefff`).

Q.13. Commentez les courbes de la figure 1 par rapport aux traits utilisés par les différents modèles.

On peut remarquer que l'ajouts de traits morphologiques apporte un gain notable en performance quelle que soit le nombre d'itération. L'ajout du lexique syntaxique permet de faire converger le modèle plus rapidement, mais n'apporte pas un gain si suffisament d'itérations sont faites.

Q.14. Le modèle 3, qui semble meilleur sur la figure 1 donnaient pourtant de moins bons résultats sur le corpus d'évaluation $V2$ dans lequel tous les noms, les adjectifs et les verbes avaient eu des lettres permutés aléatoirement. Pourquoi à votre avis ?

Le bruit dans les formes des mots font que les traits du modèles 3 deviennent inutiles et perturbent l'apprentissage.

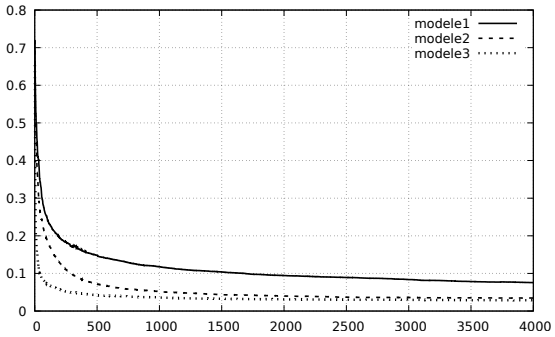

FIGURE 1 – Performance à chaque itération - classification en POS

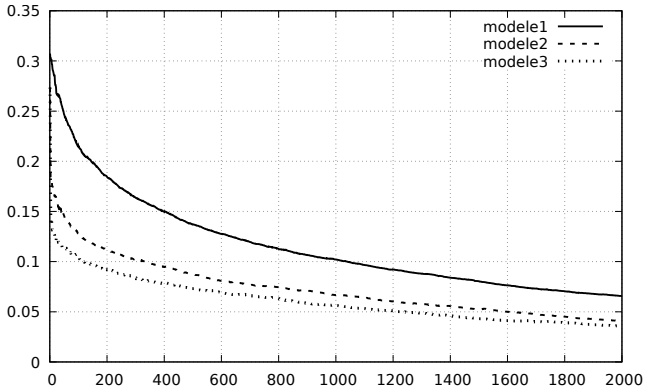

FIGURE 2 – Erreur de classification sur le corpus d'apprentissage - classification en entités nommées

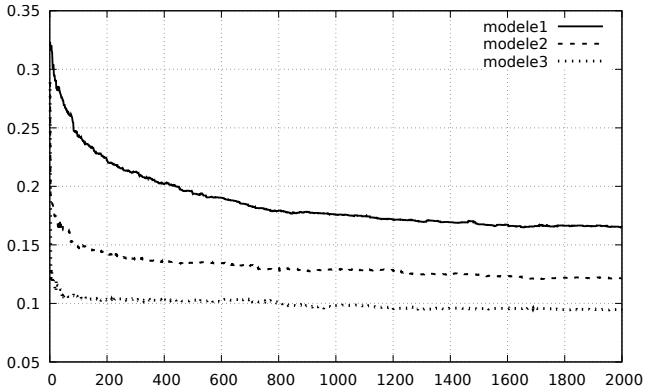

FIGURE 3 – Erreur de classification sur le corpus de développement - classification en entités nommées

Les figures 2 et 3 représentent les performances en classification en labels d'entités nommées durant les l'apprentissage de 3 modèles (2000 itérations) : le *modèle1* ne contient comme traits que le contexte immédiat du mot à étiqueter ; le *modèle2* contient en plus une information indiquant si les mots commencent ou pas par une majuscule ; enfin le modèles *modèle3* contient la liste des labels d'entités nommées possibles pour le mot à étiqueter selon un dictionnaire de référence. La figure 2 représentent les performances sur le corpus d'apprentissage à chaque itération. La figure 3 représentent les performances sur le corpus de développement à chaque itération.

Q.15. Commentez les résultats obtenus par les 3 modèles, en fonction des traits utilisés, et en commentant également les différences de comportement des courbes entre le corpus d'apprentissage et de développement.

On peut voir que l'ajout de traits sur la forme des mots (majuscules) et surtout l'ajout de dictionnaire améliorent beaucoup les résultats. Par contre le contraste entre le corpus d'apprentissage et de développement montrent que le modèle n'arrive pas à généraliser : les performances stagnent au bout de quelques itérations.

Q.16. A quoi servait dans le TP l'étape de sélection d'hypothèses par patron de POS qui était le préalable à la classification par *icsiboost* ?

A limiter le nombre d'exemples constituant le corpus donné au classifieur en entités nommées.

Q.17. Comment avez vous choisi ces patrons dans le TP évaluation ?

Q.18. Quelles, à votre avis, sont les pistes d'amélioration des modèles d'étiquetage en entité nommées que vous avez mis au point durant le TP ?