Faculté des Sciences
Aix*Marseille Université

Année universitaire 2021/2022

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 (6 pts)

Q.1. Quelles sont les similitudes et les différences entre l'analyse d'un texte par une méthode de TAL et la compilation d'un programme source, par exemple en langage C ? 1
analyse lexicale, syntaxique et semantique. Mais pas de gestion de l'ambiguite dans la compilation.

Q.2. Combien de sens différents pouvez vous trouver à la phrase : *Je vois l'homme sur la colline avec un télescope.* Quels critères peut utiliser un analyseur syntaxique automatique pour choisir une analyse particulière ? 0,5
Au moins 3. Choisir comment rattacher les rattachements prepositionnels a partir de frequences sur les paires hommes/colline ou colline/telescope, ou choisir les positions les plus probable cad les plus proches

Q.3. La plupart des modèles numériques développés pour le TAL sont basés sur de l'apprentissage supervisé. Comment est obtenu cette supervision ? Est-elle toujours facile à obtenir ? Donnez des exemples pour des tâches de TAL telles que la détection d'entités nommées, l'analyse syntaxique et la traduction automatique. 1
Supervision humaine. Difficile et chere ! ! ou de basse qualite. Pour EN, analyse syntaxique, experts, pour trad, nombreux exemples dans la vraie vie

Q.4. Quelles sont les différentes manière de représenter la notion de *mot* dans un modèle de Traitement Automatique de la Langue ? 1
dans un dictionnaire ou sous la forme d'un vecteur

Q.5. Quels sont les différences entre l'évaluation de systèmes d'étiquetage (en partie de discours ou en entités nommées par exemple) et des systèmes de traduction automatique ou de résumé de texte ? 0,5
evaluation par rapport a une reference unique pour etiquetage, references multiples pour les autres

Q.6. Comment sont évalués les systèmes de traduction automatique ou de résumé de texte ? 0,5
soit par des jugements humains, soit par des scores automatique tels que BLEU ou ROUGE qui sont au mieux correles avec le jugement humain

Q.7. Qu'est ce que la sémantique distributionnelle ? 0,5
c'est la semantique qui se base sur des proximites entre termes : le sens d'un mot, c'est son contexte d'utilisation

Q.8. Pourquoi a-t-on besoin de découper les données en 3 corpus (apprentissage, développement et test) pour développer un modèle de TAL ? Si on a un ensemble de 1000 données, combien à votre avis doit-on mettre de données dans chacun des 3 corpus ? 1
on apprend le modele sur le train, on le regle sur le dev, on l'evalue sur le test. 70,20,10

2 Analyse lexicométrique, loi de Zipf (4 pts)

Q.9. Qu'est-ce que la loi de Zipf ? 1
une loi de distribution des mots dans un texte qui dit que le ratio rang x frequence est constant

Q.10. Quelle sont ses implications pour les systèmes de Traitement Automatique du Langage ? 1
il y a un tout petit ensemble de mots qui represente la majorite des occurrences

Q.11. A quoi sert l'étape de tokenization ? 1
a segmenter un texte en unites appartenant a un dictionnaire

Q.12. Dans le TP sur l'analyse lexicométrique, vous deviez étudier le rapport entre la taille d'un corpus et la taille de son lexique (c'est à dire du nombre de mots différents le composant), puis de comparer les résultats obtenus sur un corpus contenant des transcriptions de parole et un corpus de langue écrite. A quelles conclusions étiez vous arrivé

sur ces deux questions ? 1
la courbe sur l'oral progressait moins vite, moins de mots differents

3 Etiquetage en partie de discours (POS) (5 pts)

Q.13. A quoi sert l'étape d'analyse en parties de discours (*POS tagging*) dans un processus de TAL ? 0,5
A trouver la nature des mots (nom, verbe,..)

Q.14. Quelles sont les principales ambiguïtés de cette tâche ? 0,5
certains mots peuvent etre des noms, des verbes .. beaucoup d'ambiguite sur certains mots tres communs

Q.15. Comment peut-on évaluer les performances d'un étiqueteur en parties de discours ? 1
en comparant la sequence d'etiquette predite a la reference : taux d'ettiquette correct global

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.16. Quel est à votre avis l'intérêt d'utiliser un lexique syntaxique tel que le `lefff` ? 1
A limiter l'ambiguite, permet de savoir si une forme n'est pas ambigue

Q.17. Pourquoi, malgré l'utilisation d'un lexique syntaxique, le modèle3 continue-t-il à faire des erreurs ? 1
Tous les mots ne sont pas dans le LEFFF et les formes ambigues

Q.18. Le modèle 3, qui semble meilleur sur la figure 1 donnaient pourtant de moins bons résultats que le modèle 2 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 ?
1
Car il ne peux pas utiliser les infos du lefff, toutes les formes sont inconnus, les traits du lefff deviennent du bruit

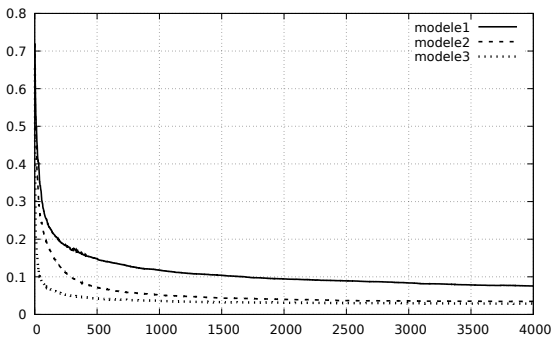

FIGURE 1 – Performance à chaque itération - classification en POS

4 Entités nommées (5 pts)

Q.19. Quelle est l'intérêt de la tâche de détection d'entités nommées ? Donnez des exemples d'applications utilisant cette tâche.
0,5
trouver les infos importantes (noms de personnes, lieux, date) dans un texte. Recherche documentaire, resume automatique, ..

Q.20. Quelles sont les deux sources d'erreur possibles dans l'évaluation d'un détecteur d'entités nommées ? 0,5
erreurs de segmentation et erreur de type

Q.21. Soit deux systèmes de détection d'entités nommées : le système A et le système B. Le système A obtient sur un corpus de test T une précision de 0.2 et un rappel de 0.8. Le système B obtient sur T une précision de 0.8 et un rappel de 0.2. Est-ce que les systèmes A et B ont la même valeur de F-score (*F1*) sur T ? Dans quel cas d'utilisation peut-on dire que A est meilleur que B ou B meilleur que A ? 1

oui meme F-score. Si on veut un systeme plus precis quand il prend une decision, c'est B, si on veut une couverture plus grande c'est A

Q.22. Quelle était la valeur de l'ambiguïté moyenne des entités nommées sur le corpus du TP (c'est à dire nombre moyen d'étiquettes différentes d'entité nommée pour une même séquence de mots) ? Cette ambiguïté était-elle dépendante de la longueur des entités ? 1
autour de 1, oui l'ambiguite decroit avec la taille

Q.23. En TP vous deviez produire la courbe ayant en abscisse les 50 patrons d'entités nommées les plus fréquents classés par fréquence et en ordonnée le nombre d'entités nommées correspondant à chaque patron dans le corpus. Que pouviez vous dire sur cette courbe ? 1
chute tres brutale juste apres les 2 premiers patrons ! !

Q.24. Quelles différences avaient vous pu observer entre l'évolution des performances entre corpus d'apprentissage et corpus de développement du classifieur icsiboost sur le tâche d'entités nommées par rapport à la tâche d'étiquetage morpho-syntaxique ? 1
sur le POS, perf equivvalente entre app et dev, pas du tout pour EN : ca baisse sur le train, mais plat sur le dev