M2 IAFA — Examen Traitement automatique du langage naturel — 2022-2023

NB:

- **Documents autorisés.**
- **Les questions sont indépendantes les unes des autres et peuvent être abordées dans l'ordre de votre choix.**
- **Estimation du temps de réponse par question : Question 1 (20 min), Question 2 (1h), Question 3 (40 min)**

---

1. Ci-dessous vous trouverez quelques phrases avec une ou plusieurs ambiguités potentielles :

   1. Je manque de pièces pour payer le café.
   2. Dans mon appartement, il y a une seule grande pièce.
   3. J'ai cassé un vase. Il y avait des pièces partout.
   4. Je suis allé voir une pièce hier avec une amie et elle était très ennuyeuse.

   (a) Identifiez les ambiguités possibles, et identifiez le niveau de langage où les ambiguités se situent (sémantique, syntaxique ou pragmatique). Donnez l'interprétation la plus plausible en justifiant.

   (b) Combien de sens de *pièce* pouvez vous identifier dans ces phrases ? Donnez un synonyme ou un hyperonyme (précisez) pour chaque sens.

   (c) Si on donne à un modèle de langue pré-entrainé, comme chatGPT, l'entrée *Quelle est l'interprétation la plus probable de la phrase <phrase>+* il continue comme suit (la phrase donnée est incluse dans la suite) : *"Il est difficile de dire quelle est l'interprétation la plus probable du texte "On a skié dans la soupe ce weekend." sans plus de contexte. Si le contexte indique que la personne se trouvait dans un endroit où il y avait de la soupe, comme dans une station de ski où il y avait un restaurant proposant de la soupe, alors l'interprétation selon laquelle la personne a skié dans un endroit où il y avait de la soupe est la plus probable. Si, en revanche, le contexte indique que la personne se trouvait dans un environnement où il n'y avait pas de soupe, alors l'interprétation selon laquelle la personne a skié dans une substance ressemblant à de la soupe est plus probable."*

   Sa réponse vous parait-elle exacte ? Qu'est-ce que cela révèle des capacités ou limites de ce genre de modèle ?

---

2. L'arrivée du navire humanitaire Ocean Viking dans le port de Marseille en novembre 2022 a suscité de nombreuses réactions sur les réseaux sociaux au sujet de la politique migratoire de la France. Pour analyser l'impact de ce flux migratoire sur la population et mesurer les stéréotypes que cela peut déclencher (par exemple, "les migrants sont violents", "les migrants vont profiter des aides sociales"), on se propose de développer un système de détection automatique de messages de haine pour le français[1]. On se focalisera sur le réseaux social Twitter.

   I. On veut construire un classifieur qui est capable de reconnaître automatiquement si un message est haineux ou non.

   (a) De quelles sortes de données d'entraînement a-t-on besoin ?

[1] Un stéréotype n'est pas forcément haineux et vice-versa.

(b) Détaillez brièvement la démarche à suivre pour collecter ces données d'entraînement ? Détaillez en particulier les précautions à prendre lors de cette collecte afin d'éviter les biais d'apprentissage.

(c) Détaillez brièvement la démarche à suivre pour construire ces données d'entraînement ? Comment mesurer leur qualité ?

II. Après les phases de collecte et de construction, on se retrouve avec corpus de 10K tweets avec une proportion de messages haineux de 20%.

(a) Pensez-vous utile d'équilibrer les données. Si oui, expliquez comment procéder à cette équilibrage. Si non, justifiez.

(b) Pensez-vous utile de pré-traiter les données. Si oui, quelles sont les pré-traitements qui vous semblent les plus pertinents. Si non, justifiez.

(c) On souhaite rendre les données disponibles à la communauté dans un repo type GitHub ou Zenodo. Expliquez comment doit-on procéder et les précaution à prendre.

II. On se propose maintenant d'entraîner trois classifieurs binaires sur nos données : (M1) Un premier classifieur baseline bag of words, (M2) un second classifieur à base de traits, (M3) un modèle neuronal.

(a) Expliquez comment fonctionne la baseline (M1).

(b) Expliquez comment fonctionne le modèle (M2). Quels traits vous semblent les plus pertinents ? Quel algorithme d'apprentissage utiliseriez-vous ?

(c) Quelle architecture neuronale utiliseriez-vous pour battre (M1) et (M2) ? Expliquez les entrées de votre modèle (M3), les sorties qu'il doit prédire ainsi que le fonctionnement de cette architecture. Précisez la fonction loss.

(d) Quelle métrique d'évaluation est la plus adaptée pour évaluer vos modèles ? Justifiez.

(e) L'analyse des performances du modèle (M2) et celui proposé en (M3) montre que les scores diffèrent de 0.03% en faveur de (M3). Peut-on dire que le modèle (M3) est le plus performant ? Justifiez.

(f) La dernière étape est l'évaluation qualitative de vos modèles. Expliquez en quoi consiste cette évaluation.

(g) Si on voulait une idée des informations qu'utilise le modèle (M3) pour prendre sa décision, quelle(s) méthode(s) pourrait-on utiliser ? Quel type d'information donneraient-elles ?

III. On veut améliorer les performances de ce classifieur de sorte qu'il soit également capable de mieux appréhender le contexte d'énonciation des messages, c'est à dire prendre en compte des informations linguistiques pertinentes mais non présentes dans le message à classer. Par exemple la classification des messages[2] "*On doit tous les renvoyer chez eux <URL>*", "*On doit tous les renvoyer chez eux <IMAGE>*", "*On doit tous les renvoyer chez eux*" est rendu difficile par l'absence de contexte.

(a) Le classifieur proposé en (II) est-il adapté ? Justifiez votre réponse.

(b) Proposez une solution pour adapter le classifieur initial pour prendre en compte un contexte d'énonciation donné. *Attention, il n'est pas demandé une solution pour TOUS les contextes d'énonciation. Fixer un type de contexte et proposer une solution par rapport à ce contexte.*

---

3. On va s'inspirer du texte suivant pour un problème d'extraction d'information, où on veut repérer des noms de modèles d'ordinateur dans des documents.

   Après avoir servi pendant la Seconde Guerre mondiale, Grace Hopper est devenue chargée de recherche à Harvard où elle a travaillé sur les ordinateurs Mark II et Mark III avant de passer dans le secteur privé.

[2] Dans ces exemples, <URL> et <IMAGE> désignent des liens externes vers des images, articles de presse, etc.

En 1952, alors qu'elle supervisait la programmation de l'ordinateur UNIVAC chez Remington Rand, elle et son équipe ont créé le premier compilateur, un programme qui convertit les instructions du langage en code, en langage informatique, afin que les informations puissent être lues et exécutées par un ordinateur.

On attribue également à Hopper la popularisation des termes «bug informatique» et «débogage». Ces termes sont nés d'un incident survenu pendant son séjour à Harvard, lorsqu'on a découvert qu'une mite avait court-circuité l'ordinateur d'Harvard Mark II.

(a) Dans le cadre d'une approche non supervisée, quel(s) patron(s) lexical(ux) d'extraction pourrait-on utiliser ? Evaluer la précision et le rappel des patrons proposés sur le texte exemple. Pouvez vous trouver un patron avec un rappel de 1 sur le texte choisi ?

(b) Comment pourrait-on trouver de nouveaux patrons ?

(c) Si on veut développer un modèle supervisé pour le même problème, quel type d'architecture neuronale pourrait-on utiliser ? Préciser ce que serait une instance du problème, les entrées correspondant et les sorties que doit prédire le modèle, en prenant comme exemple une phrase (ou partie de phrase) pertinente du texte précédent.

(d) Si on donne la phrase suivante en entrée de BERT, « En 1952, alors que Grace Hopper supervisait la programmation de l'ordinateur UNIVAC chez Remington Rand, [MASK] et son équipe ont créé le premier compilateur. » Les trois mots les plus probables sont "Hopper" (98%), "lui" (0.4%), "elle" (0.1%). Pouvez vous expliquer pourquoi ?