Université de Montréal

IFT6285 (TALN) — Projet 1
Fouille dans l'anthologie d'ACL pour identifier des tâches de classification sous-exploitées et pour les essayer

Contact :
Philippe Langlais
RALI/DIRO
Université de Montréal

+1 514 343 61 11 ext: 47494
felipe@iro.umontreal.ca
http://www.iro.umontreal.ca/~felipe/

dernière compilation : 16 octobre 2025 (10:53)

Contexte

L'anthologie de l'Association for Computational Linguistics (ACL) répertorie des articles scientifiques publiés dans des tribunes scientifiques diverses. Une API permet d'interroger cette anthologie dans le confort de Python.

Dans ce projet votre but est d'une part de caractériser l'évolution de la problématique de la classification dans les articles répertoriés dans l'anthologie et d'identifier des tâches de classification qui ne sont pas populaires, afin d'essayer sur ces tâches différentes technologies.

À faire

1. Vous devez faire une **cartographie** des travaux publiés dans conférences ACL (incluant NAACL et EACL), CoNLL, EMNLP, COLING et LREC, ainsi que les Findings concernant des *benchmarks* ou jeux de données qui ont trait à la classification. Des informations comme le nombre de benchmarks proposés, leur caractérisation en type (classification binaire versus multi-classe ou multi-labels) ou en domaine (détection de sentiments, de polarité), les performances de la meilleure approche testée sur ces benchmarks ou encore le nombre de citations au jeux de données peuvent être étudiées et rapportées.

   Ceci peut se faire en plusieurs temps : 1) récupérer dans un format de votre choix les articles des conférences d'intérêt ; 2) rechercher les articles parlant de classification, 3) colliger les informations pertinentes, possiblement manuellement. Par exemple, l'article 2023.acl-long.461 propose un benchmark (MAD-TSC) qui devrait être comptabilisé dans votre portrait.

2. Dans un deuxième temps, vous allez prendre au moins **deux tâches de classification** que vous avez repérées et allez essayer différentes approches pour les résoudre. Vous n'êtes pas contraint sur les technologies que vous pouvez développer, mais le but est de vérifier ce que différentes approches apportent. Je vous conseille d'essayer des approches simples à l'aide d'une boite à outils comme scikit-learn et de comparer cela à des approches comme interroger un grand modèle de langue ou encore affiner un modèle de type BERT.

Votre but n'est pas tant de battre l'état de l'art que d'observer ce que différentes approches font, d'étudier leur limites et leurs éventuelles complémentarités. Vous pouvez bien sûr vous comparer aux approches testées dans l'article décrivant le jeu de données. Cela peut se faire en reprenant les résultats décrits dans l'article ou en lançant le code si disponible. L'un de vos buts est également d'écrire un code générique afin de minimiser votre intervention pour traiter une tâche de classification particulière.

La seule contrainte concernant les tâches que vous allez étudier est qu'elle ne soient pas populaires : le but n'est pas ici de lancer des expériences sur des tâches comme GLUE (article cité 9352 fois en date de rédaction de ce sujet) ou SUPER-GLUE (article cité 2923 fois). En fait, si un article est cité plus de 200 fois, vous ne pouvez pas étudier le jeux de données associé.

3. Vous devez produire un **rapport d'au plus 8 pages** (format pdf, en français ou en anglais) dans lesquelles vous livrerez votre cartographie des articles portant sur la classification et décrirez vos expériences de classification. Une bibliographie comptant dans ces 8 pages listera à minima les références aux articles introduisant les jeux de données que vous avez étudiés. Vous pouvez ajouter des annexes qui ne seront pas comptabilisées dans les 8 pages, mais le texte utile doit faire partie des 8 pages autorisées (vous pouvez mentionner des informations secondaires dans l'annexe). Je vous suggère d'utiliser le format ACL (double colonne) pour produire votre rapport (ou tout autre style similaire). Ce rapport a pour objectif de mettre en valeur la connaissance que votre travail permet d'engendrer. Ce rapport sera lu comme un article scientifique.

Notation

La notation n'est pas corrélée à la performance de vos approches, mais à la **curiosité** que vous développerez et à votre esprit d'**analyse**. La clarté et l'information de vos rapports sont deux critères très importants. Il est a priori inutile de mettre du code dans votre rapport (vous pouvez en mettre dans les annexes). Le volet cartographie est aussi important que le volet classification.

Remise

Le travail est à faire seul ou en binôme. La remise est à faire sur Studium sous le libellé `projet1`. [1] Vous devez remettre dans un fichier `<nom(s)>.tar.gz` votre rapport (`rapport-<nom(s)>.pdf`) où `<nom(s)>` contient le nom/logging DIRO des personnes impliquées par le projet (ex : `felipe-petrov`.). La remise est à faire au plus tard **vendredi 14 novembre à 23h59**.

1. Une seul remise par binôme.