Etudes expérimentales pour l'interprétabilité des mesures d'accord inter-annotateur

Responsables

Jean-Yves Antoine (Jean-Yves.Antoine@univ-tours.fr) LIFAT U. Tours
Anaïs Halftermeyer (anais.halftermeyer@univ-orleans.fr) LIFO, U. Orléans
Jeanne Villaneau (jeanne.villaneau@univ-ubs.fr) IRISA, Lorient
Farida Saïd (Farida.Saïd@univ-ubs.fr) IRISA, Lorient

Résumé

Etude expérimentale de la validité d'une correspondance entre mesures d'accord inter-annotateur et stabilité d'une annotation de référence, par génération aléatoire de données bruitées et observation du comportement de ces métriques.

Contexte scientifique

Le Laboratoire d'Informatique de l'Université de Tours (LIFAT) propose un stage en collaboration avec le laboratoire LIFO de l'Université d'Orléans et l'IRISA Lorient (INRIA), dans le cadre d'un financement de la fédération ICVL (Informatique Centre Val de Loire). Ce stage fait suite à une collaboration déjà initiée par le passé sur l'étude de la validité des métriques d'accord inter-annotateur.

Le machine learning (techniques statistiques d'apprentissage automatique) et le deep learning (réseaux de neurones artificiels) font souvent appel, pour leur entrainement, à des données qui ont été obtenues par annotation manuelle réalisées par la foule (crowdsourcing) ou par des experts humains. Par exemple, pour les applications de suivi d'opinion (*opinion mining*) ou d'analyse de sentiments (*sentiment analysis*) sur les réseaux, on recourt à des lexiques émotionnels pour lesquels un ensemble d'annotateurs a associé à chaque mot une valeur émotive (allant de *très négatif* à *très positif*). Les techniques de contrôle d'accès par identification d'images telles que reCAPTCHA vous font également jouer le rôle d'annotateur de données, sans que vous en soyez avertis. Dans tous les cas de figure, il est donc demandé à un pool d'annotateur de classifier un ensemble d'observables. On construit alors un corpus dit de référence à partir d'un vote majoritaire sur ces annotations

Le processus d'annotation étant le plus souvent subjectif, la qualité des données est évaluée par des mesures de fiabilités reposant sur une mesure d'accord inter-annotateurs : plus une foule d'annotateur s'accorde sur une annotation de référence, plus celle-ci a des chances a priori d'être fiable. De nombreuses métriques d'accord ont été proposées qui diffèrent par leur manière d'estimer l'accord au hasard entre les annotateurs. En effet, avoir un bon accord brut n'a pas la même signification suivant que la chance d'arriver à cet accord était grande ou faible.

Une des limites principales de ces mesures d'accord (outre leur sensibilité à certains biais statistiques bien étudiée dans la littérature) est leur manque d'interprétabilité. Les mesures d'accord utilisées n'ont en effet jamais fait l'objet d'études expérimentales fouillées permettant de définir des échelles d'interprétation objectives et parlantes. Pour l'heure, la

communauté scientifique en est réduite à des arguments d'autorités du type « une valeur de 0,64 est acceptable », sachant que les différents scientifiques qui se sont aventurés sur le sujet n'ont pas donné d'avis cohérents pour des valeurs d'accord identiques.

Dans le cadre de nos recherches actuelles, nous avons proposé d'interpréter les mesures d'accord d'une manière objective et intelligible en reliant ces dernières avec le taux de stabilité de l'annotation de référence. Cette stabilité est estimée à partir du taux de variation moyen de la référence, c'est-à-dire la probabilité moyenne qu'a une annotation de référence obtenue avec N annotateur de changer si l'on prenait N nouveaux annotateurs. Ainsi, on pourrait relier une valeur d'accord donné à une estimation, facilement interprétable, du degré de stabilité de l'annotation de référence.

Nos premiers travaux sur le sujet ont montré la pertinence de notre démarche, et ont déjà l'objet d'une publication dans une revue scientifique (https://hal.archives-ouvertes.fr/hal-02375240). Si ces études ont montré qu'il existait une relation entre valeur d'accord inter-annotateur et taux de variation de la référence, elles ne nous pas encore permis de caractériser proprement les biais statistiques qui caractérisaient cette dernière. Deux facteurs d'influence ont en particulier été identifiés, qui doivent être étudiés plus précisément :

- La distribution des désaccords entre annotateurs suivant les items à annoter
- La distribution des désaccords entre annotateurs suivant la classe d'annotation considérée

Le sujet proposé consiste à étudier plus finement la corrélation entre une mesure d'accord, le α de Krippendorff, et le taux de stabilité de l'annotation de référence, en travaillant sur des données réelles et synthétiques générées spécifiquement afin de fait varier artificiellement l'accord entre les annotateurs. Ce stage consistera à reprendre le code d'un outil permettant de calculer ces métriques sur un jeu de données fourni, à récupérer de nouvelles données annotées réelles et enfin à étudier expérimentalement le comportement de la métrique sur les données, afin d'en arriver à une meilleure interprétation de la relation entre α et taux de variation. Suivant les résultats et l'envie de la personne recrutée, il pourra être envisagé de participer à une publication présentant ce travail. Il est à noter que ce travail, de portée générique (i.e. concernant tout type d'annotation), se focalisera avant tout sur des applications dans le domaine du Traitement Automatiques des Langues Naturelles (TALN).

Travail à réaliser (estimation temporelle en semaines)

- **Phase 1 : prise en main de l'outil de calcul d'accord inter-annotateur et découverte de la problématique (S0 – S0+2)** – Un outil permettant le calcul de α et la génération de données simulées a déjà été développé en C++ par le passé par Jeanne Villaneau.
- **Phase 1 : récupération de données annotées libres et mise en format exploitable (S0 – S0+6)** – Les études que nous avons menées jusqu'à présent manquent de représentativité, ne portant pour l'heure que sur 5 corpus annotés différents. Cette étape consistera à récupérer d'autres données libres annotées, issues du TALN mais pouvant éventuellement être ouvertes à d'autres domaines, et de modifier leur représentation pour la rendre utilisable par notre système.
- **Phase 3 : évaluation expérimentale du comportement des mesures d'accord inter-annotateur (S0+6 – S0+12)** – Cette étape consistera à utiliser l'outil développé et les corpus récupérer pour approfondir nos études actuelles sur la relation entre mesure d'accord et taux de variation. Cette dernière phase revêt une dimension recherche

affirmée et mettra en jeu des notions statistiques de base. Elle pourra donner lieu, suivant les résultats expérimentaux obtenus, à des développements informatiques permettant d'intégrer de nouveaux modèles statistiques de représentation ou d'analyse des données.

Résultats attendus

- Adaptation éventuelle du système de calcul d'accord et de génération aléatoire d'annotations
- Etablissement d'un modèle clair d'interprétation des mesures d'accord (□ de Krippendorff) en fonction de l'estimation du taux de variation moyen de l'annotation de référence.

Profil recherché

Ce stage demande des capacités de développement objet relevant d'un niveau d'études de **licence troisième année en informatique** ou de **fin de première année d'ingénieur en informatique**. Mais avant tout, on attend de la personne recrutée qu'elle présente un intérêt marqué pour la recherche, qu'elle ait une autonomie et un sens critique développés, et qu'elle ne soit pas rétive à considérer quelques notions mathématiques de base en matière d'estimation statistique de l'accord par chance.

Ce stage de découverte est donc proposé à des étudiants qui disposeraient d'un excellent niveau académique, d'une curiosité scientifique affirmée et qui réfléchiraient à une orientation professionnelle future dans le domaine de la recherche.

Date et lieu de stage

La personne recrutée travaillera au sein du laboratoire LIFAT (antenne universitaire de Blois) où elle s'intégrer dans l'équipe BDTLN (http://li.univ-tours.fr/equipes/equipe-bdtln-198022.kjsp). Elle travaillera en collaboration constante avec Anaïs Halftermeyer, de l'équipe Contraintes et Apprentissage (http://www.univ-orleans.fr/lifo/equipes/CA/) du LIFO (U. Orléans) ainsi qu'avec Jeanne Villaneau et Farida Saïd, de l'équipe EXPRESSION (https://www-expression.irisa.fr/research/text/) de l'IRISA). En outre, un séminaire de recherche mensuel autour de la langue naturelle (RITUEL) est organisé entre les centres de recherche des universités de Tours (LI) et Orléans (LIFO, LLL). La personne recrutée sera invitée à y participer si elle le souhaite.

Durée et période de stage – La durée du stage sera de 3 mois. Début de stage à négocier avec la personne sélectionnée (mi-avril 2021 au plus tard).

Rémunération

La personne recrutée recevra une gratification mensuelle correspondant à la réglementation, à savoir 15% du plafond horaire de la sécurité sociale. A titre d'exemple, cette gratification représente un montant de 554 € pour un mois avec 22 jours ouvrés, et 600,60 € pour un mois avec seulement 20 jours ouvrés (jours fériés, par exemple). Pourra également se rajouter une indemnité de transports en commun correspondant à 50% d'un abonnement mensuel étudiant.

La personne recrutée participera aux réunions de l'équipe projet. Les frais de mission induits par ces déplacements seront remboursés.

Contact – Dépôts de candidature

Dépôt des candidatures par courrier électronique auprès de Jean-Yves Antoine, Anaïs Lefeuvre-Halftermeyer, Jeanne Villaneau et Farida Saïd, avant le **4 janvier 2021**, délai de rigueur. Merci de déposer :

- Un CV détaillé de vos activités passées
- Une lettre de motivation
- Vos relevés de notes des deux dernières années d'études

Le cas échéant, un développement Java et/ou une lecture critique d'article scientifique pourront être demandé pour la sélection.

Références

Antoine J.-Y., Villaneau J., Lefeuvre A. (2014) Weighted Krippendorff's alpha is a more reliable metrics for multi-coders ordinal annotations: experimental studies on emotion, opinion and coreference annotation. Proc. 14th *Conference of the European Chapter of the Association of Computational Linguistics, EACL'2014*, Gothenburg, Suède [http://hal.archives-ouvertes.fr/hal-01001811]

Arstein R., Poesio M. (2008). Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*. 34(4):555-596. [https://www.cairn.info/revue-langue-francaise-2015-4-page-5.htm]

Brégéon D., Villaneau J, Antoine J.Y., Halftermeyer A. (2019) Redonner du sens à l'accord inter-annotateurs : vers une interprétation des mesures d'accord en termes de reproductibilité de l'annotation. Revue *TAL*, ATALA, 60(2). (https://hal.archives-ouvertes.fr/hal-02375240)

Mathet, Y. and Widlöcher, A. (2016). Évaluation des annotations : ses principes et ses pièges. Revue *T.A.L.*, 52(2):73–98 [https://www.atala.org/Evaluation-des-annotations-ses]