Sujet de thèse 2018

Evaluation des annotations par des mesures d'accord inter-annotateurs : le cas des flux (texte, audio, vidéo)

Encadrement : Yann Mathet (HDR), Antoine Widlöcher
Laboratoire : GREYC CNRS UMR 6072, Université de Caen Normandie
Équipes de rattachement : HULTECH, CODAG
Contact : yann.mathet@unicaen.fr

Date limite de candidature : 14 mai 2018
Début de la thèse : septembre ou octobre 2018
Durée : 3 ans
Allocation ministérielle

La linguistique computationnelle et le traitement automatique des langes ont besoin de s'appuyer sur des données annotées pour développer et évaluer les modèles et les traitements.

Dans ce contexte, afin d'établir des annotations de référence (gold standard), on a souvent recours à l'annotation manuelle multiple : il s'agit de soumettre les mêmes données d'entrée (par exemple un texte) à plusieurs annotateurs humains indépendants, puis de comparer leurs annotations. L'hypothèse est que si leur degré d'accord est suffisant, il est alors possible d'établir un gold standard à partir de leurs productions (via un arbitrage, différentes stratégies étant possibles). Le rôle des mesures d'accord inter-annotateurs est justement de quantifier ce degré de consensus entre les annotateurs.

Les mesures bien connues et utilisées depuis des années, telles que le fameux κ de Cohen, cf. [Cohen, 1968], ne remplissent pourtant pas cet objectif, dans la plupart des cas, dans le domaine du TAL, et leur utilisation conduit à des valeurs largement biaisées. En effet, notre objet d'étude étant par nature un flux de données (un "continuum"), que ce soit via un texte, un enregistrement audio, ou une vidéo, les annotations auxquelles nous sommes confrontés sont de type "unitizing", c'est-à-dire que les annotateurs doivent placer librement des unités de taille variable où ils le souhaitent sur le flux, alors que les mesures classiques prennent en entrée des items prédéfinis (d'où les importants biais qui résultent de l'utilisation détournée qui en est faite en TAL). Seuls les coefficients α de Krippendorff dédiés à l'unitizing tentent de proposer des solutions adaptées, mais ils ne correspondent pas idéalement aux besoins du TAL, comme nous l'avons montré dans [Mathet et al., 2012].

Les équipes Hultech et Codag du GREYC ont entamé un travail de longue haleine depuis 2010 afin de proposer des mesures adaptées à l'unitizing. En particulier, une nouvelle famille de mesures d'accord, les γ, ont récemment vu le jour, et ont reçu une reconnaissance au niveau international, comme en témoignent les deux publications dans la revue Computational Linguistics [Mathet et al., 2015] et [Mathet, 2017]. Cette expertise s'est aussi traduite par une collaboration avec Klaus Krippendorff, spécialiste reconnu, afin d'améliorer et d'implémenter ses mesures d'accord, cf. [Krippendorff et al., 2016], et par la publication d'un article méthodologique dans la revue TAL, cf. [Mathet and Widlöcher, 2016], soit un total de 4 acticles dans des revues internationales entre 2015 et 2017.

Cette dynamique offre un cadre idéal à un travail de recherche en thèse, d'autant que beaucoup de questions restent ouvertes dans le domaine des mesures d'accord en général, et de l'unitizing en particulier. Ce travail de recherche s'attachera, de façon non limitative, à :

- Etendre les mesures γ, qui concernent actuellement les unités, aux relations entre ces dernières. Il s'agirait une extension très attendue dans des domaines tels que l'analyse des relations du discours ou l'étude des chaînes de référence.
- Les mesures γ sont actuellement génériques, c'est-à-dire traitent de la même façon tout

continuum. Cependant, différentes tâches d'annotation nécessiteraient des ajustement spécifiques. Par exemple les désaccords sur la position des bornes des unités n'ont pas la même importance selon la tâche d'annotation.

— La prise en compte des attributs-valeurs, étant donné que les mesures actuelles, y-compris les γ, prennent en compte les catégories annotées, mais jamais les traits qui leur sont parfois associés.

— La question de la part de chance dans les mesures d'accord. Les mesures d'accord étant (pour la plupart) corrigées par la chance, mais les stratégies pour estimer cette dernière étant multiples, il s'agira de les comparer et de proposer des améliorations.

Ce sujet de thèse est candidat à une allocation doctorale ministérielle parmi d'autres sujets (audition fin mai, éventuellement possible en visio-conférence).

La candidature, au format PDF (en un seul fichier), pourra comporter un CV, une lettre de motivation, une liste de publications, d'éventuelles lettres de recommandation. Nous pouvons répondre à vos questions par mail avant votre éventuelle candidature.

Références

[Cohen, 1968] Cohen, J. (1968). Weighted kappa : Nominal scale agreement with provision for scaled disagreement or partial credit. In *Psychological Bulletin*, volume 70, pages 213–220.

[Krippendorff et al., 2016] Krippendorff, K., Mathet, Y., Bouvry, S., and Widlöcher, A. (2016). On the reliability of unitizing textual continua : Further developments. *Quality and Quantity Journal*, 50(6) :2347—2364.

[Mathet, 2017] Mathet, Y. (2017). The agreement measure gamma-cat (γ_{cat}), a complement to gamma focused on categorization of a continuum. *Computational Linguistics*, 43(3) :661–681.

[Mathet et al., 2012] Mathet, Y., Widlöcher, A., Fort, K., François, C., Galibert, O., Grouin, C., Kahn, J., Rosset, S., and Zweigenbaum, P. (2012). Manual Corpus Annotation : Giving Meaning to the Evaluation Metrics. In *Proceedings of the International Conference on Computational Linguistics (COLING 2012)*, pages 809–818, Mumbaï, Inde.

[Mathet et al., 2015] Mathet, Y., Widlöcher, A., and Métivier, J.-P. (2015). The unified and holistic method gamma (γ) for inter-annotator agreement measure and alignment. *Computational Linguistics*, 41(3) :437–479.

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