6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Évaluation des annotations par des mesures d'accord inter-annotateurs

Anaëlle Baledent
Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC, 14000 Caen, France
anaelle.baledent@unicaen.fr

RÉSUMÉ

Nous présentons dans ce descriptif notre sujet de thèse portant sur l'évaluation des annotations par des mesures d'accord inter-annotateurs. Ces mesures permettent d'établir, à partir d'annotations manuelles multiples, des corpus de référence, dont leur constitution est un enjeu pour le Traitement Automatique des Langues. L'objectif de cette thèse est notamment de conseiller et d'outiller les chercheurs sur les mesures d'accord inter-annotateurs, afin d'améliorer la qualité des annotations de référence.

ABSTRACT

Annotations evaluation by inter-annotator agreement measures

In this description, we present our thesis subject on annotations evaluation by inter-annotator agreement measures. These measurements make it possible to establish, from multiple manual annotations, gold standard, whose constitution is an issue for Natural Languages Processing. The aim of this thesis is notably to advise and equip researchers on inter-annotator agreement measures, in order to improve the quality of gold standard.

MOTS-CLÉS : accord inter-annotateur, annotation de référence, évaluation d'annotations.

KEYWORDS: inter-annotator agreement, gold standard, annotation evaluation.

1 L'évaluation des annotations manuelles multiples

Notre thèse, dirigée par Yann MATHET et Antoine WIDLÖCHER au sein du GREYC à Caen, s'inscrit dans le contexte de la création des ressources langagières pour le Traitement Automatique des Langues. Nous nous intéressons aux données annotées et à l'établissement d'annotations de référence. Pour établir un corpus de référence, on a souvent recours à l'annotation manuelle multiple : on soumet les mêmes données à plusieurs annotateurs humains, puis on compare leurs annotations en s'appuyant sur des mesures d'accord inter-annotateurs. Ces mesures permettent de quantifier le degré de consensus des annotateurs. Si ce degré d'accord est jugé satisfaisant, une référence est établie à partir des annotations.

Les mesures les plus connues, comme κ (Cohen, 1968), sont adaptées pour une annotation de type catégorisation (assigner une catégorie à une occurrence prédéfinie), mais elles ne conviennent pas pour la segmentation d'un continuum, où l'annotateur doit délimiter les bornes des occurrences en plus de catégoriser ces dernières (par exemple les structures multi-échelles annotées lors du projet ANNODIS (Colléter *et al.*, 2012)). Dans ce cas-là, des mesures telles que les α dédiés à l'unitizing (Krippendorff,

1980) ou γ (Mathet *et al.*, 2015) sont préconisées pour mesurer l'accord inter-annotateurs : en plus de la catégorisation, elles prennent en compte les problèmes d'alignement des unités (positions des bornes non correspondantes, enchâssement et/ou superposition de deux unités, etc.).

Utiliser les mesures adaptées permet d'avoir des valeurs plus pertinentes concernant l'évaluation des annotations multiples et pourrait faciliter ou améliorer la constitution de corpus de référence. Or leur qualité est d'autant plus importante que de ces *gold standard* découle la fabrication d'autres outils du TAL. Par exemple, (Manning, 2011) analyse les erreurs d'un étiqueteur morpho-syntaxique et estime que plus de 40% des erreurs sont dues à une mauvaise référence (erronée ou manquante de consistance). Mais nous observons une méconnaissance des mesures d'accord et les mauvaises utilisations biaisent l'établissement d'annotations de référence, comme démontré dans (Mathet *et al.*, 2015).

2 De la mesure d'accord à la compréhension fine des désaccords

Le principal enjeu de cette thèse est d'outiller les responsables de campagnes d'annotation, et plus généralement les chercheurs, pour l'analyse des mesures d'accord. Pour ce faire, il convient dans un premier temps de dresser une vue d'ensemble des campagnes d'annotation en TAL et d'en dégager les pratiques et les méthodologies, ainsi que cerner les manques et méconnaissances, en se focalisant sur l'utilisation des mesures d'accord. Nous nous focaliserons principalement sur des campagnes où les corpus sont multi-annotés, tels que ANNODIS, ANCOR (Muzerelle *et al.*, 2014) ou le corpus émotion produit par (Le Tallec *et al.*, 2011).

Une fois cet état de l'art établi, il s'agira ensuite d'affiner, voire créer, des mesures d'accord en prenant en compte les types de données et la tâche d'annotation. S'il existe certaines métriques d'évaluation prenant en compte les relations (par exemple les mesures UAS et LAS (Kübler *et al.*, 2009) pour les analyseurs syntaxiques), nous aimerions généraliser le principe de ces mesures à d'autres types de données. Il en va de même pour des mesures prenant en compte les attributs et les valeurs. En ce sens, les chaînes de coréférence semblent être un excellent terrain d'expérimentation sur lequel nous concentrerons nos efforts. En effet, ces chaînes reposent autant sur des mécanismes de segmentation que sur des structures relationnelles. Là encore, le projet ANCOR nous permettra de mener nos investigations.

À terme, notre travail devra aussi formuler des recommandations selon les campagnes. Elles s'appuieront sur des méthodes d'évaluations et des observations liées aux types de données et aux tâches d'annotation. Elles pourront être réalisées dès les premières annotations produites et être implémentées dans un outil. Cela permettra aux responsables des campagnes de mieux cibler les causes du désaccord en identifiant les zones et les configurations dans lesquelles le désaccord émerge en priorité, et ainsi améliorer les consignes d'annotation en conséquence. Ce pan de notre travail a un double objectif : d'une part, améliorer et faciliter l'élaboration de corpus de référence, et d'autre part mieux comprendre l'accord inter-annotateurs, comme récemment préconisé par (Bregeon *et al.*, 2019).

Références

BREGEON D., ANTOINE J.-Y., VILLANEAU J. & LEFEUVRE-HALFTERMEYER A. (2019). Redonner du sens à l'accord interannotateurs : vers une interprétation des mesures d'accord en termes

de reproductibilité de l'annotation.

COHEN J. A. (1968). Weighted kappa : nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological bulletin*, **70 4**, 213–220.

COLLÉTER M., FABRE C., HO-DAC L.-M., PÉRY-WOODLEY M.-P., REBEYROLLE J. & TANGUY L. (2012). *La ressource ANNODIS multi-échelle : guide d'annotation et bonus*. Rapport interne. HAL : hal-00983076.

KRIPPENDORFF K. (1980). *Content Analysis : An Introduction to Methodology*. Beverly Hills, CA : Sage Publications, Inc.

KÜBLER S., MCDONALD R., NIVRE J. & HIRST G. (2009). *Dependency Parsing*. Morgan and Claypool Publishers.

LE TALLEC M., ANTOINE J.-Y., VILLANEAU J. & DUHAUT D. (2011). Affective Interaction with a Companion Robot for Hospitalized Children : a Linguistically based Model for Emotion Detection. In *5th Language and Technology Conference (LTC'2011)*, p. 6 pages, Poznan, Poland. 6 pages, HAL : hal-00664618.

MANNING C. D. (2011). Part-of-speech tagging from 97% to 100% : Is it time for some linguistics ? In A. F. GELBUKH, Éd., *Computational Linguistics and Intelligent Text Processing*, p. 171–189, Berlin, Heidelberg : Springer Berlin Heidelberg.

MATHET Y., WIDLÖCHER A. & MÉTIVIER J.-P. (2015). The unified and holistic method gamma (γ) for inter-annotator agreement measure and alignment. **41**(3), 437–479. DOI : 10.1162/COLI_a_00227.

MUZERELLE J., LEFEUVRE A., SCHANG E., ANTOINE J.-Y., PELLETIER A., MAUREL D., ESHKOL I. & VILLANEAU J. (2014). ANCOR_Centre, a Large Free Spoken French Coreference Corpus : description of the Resource and Reliability Measures. In ELRA, Éd., *LREC'2014, 9th Language Resources and Evaluation Conference.*, p. 843–847, Reyjavik, Iceland. HAL : hal-01075679.