Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles
Lille, France, 28 juin au 2 juillet 2021

Extraction d'arguments basée sur les transformateurs pour des applications dans le domaine de la santé

Tobias Mayer[1] Elena Cabrio[1] Serena Villata[1]
(1) Université Côte d'Azur, CNRS, Inria, I3S, France
tmayer@i3s.unice.fr, {elena.cabrio,serena.villata}@univ.cotedazur.fr

RÉSUMÉ

Nous présentons des résumés en français et en anglais de l'article (Mayer *et al.*, 2020) présenté à la conférence 24th European Conference on Artificial Intelligence (ECAI-2020) en 2020.

ABSTRACT

Transformer-based Argument Mining for Healthcare Applications

We present French and English abstracts of the article (Mayer *et al.*, 2020) that was presented at the 24th European Conference on Artificial Intelligence (ECAI-2020).

MOTS-CLÉS : extraction d'arguments, essais contrôlés randomisés, PICO, prise de décision fondée sur des données probantes.

KEYWORDS: argument mining, randomized controlled trials, PICO, evidence-based decision making.

1 Résumé en français

L'extraction d'arguments (AM) vise généralement à identifier les composants argumentatifs dans le texte et à prédire les relations entre eux. La prise de décision basée sur les preuves (Hunter & Williams, 2012; Craven *et al.*, 2012; Longo & Hederman, 2013; Qassas *et al.*, 2015) en santé numerique vise à soutenir les cliniciens dans leur processus de délibération pour établir le meilleur plan d'action pour le cas en cours d'évaluation. Cependant, malgré son utilisation naturelle dans les applications de santé, seules quelques approches d'AM ont ete éppliqués à ce type de texte (Green, 2014; Mayer *et al.*, 2018, 2019), et leur contribution se limite à la détection des composants argumentatifs, sans prendre en compte la prédiction des relations entre eux. De plus, aucun grand ensemble de données annotées pour l'AM n'est disponible pour le domaine de la santé. Dans cet article (Mayer *et al.*, 2020), nous avons répondu à la question de recherche suivante : *comment définir un pipeline complet d'AM pour les essais cliniques ?* Pour répondre à cette question, nous proposons une approche basee sur des *transformers* bidirectionnels combinée à différents réseaux neuronaux pour la détection de composants argumentatifs et la prédiction de relations dans les essais cliniques, et nous évaluons cette approche sur un nouveau corpus de 659 résumés de la base de données MEDLINE. En particulier, nous avons étendu un jeu de données existant en annotant 500 résumés d'essais de MEDLINE, conduisant à 4198 composants argumentatifs et 2601 relations sur différentes maladies. En suivant les lignes directrices pour l'annotation des composants d'arguments dans les essais cliniques de (Trenta *et al.*, 2015), deux annotateurs ayant une formation en linguistique ont effectué l'annotation des 500 résumés sur le néoplasme. L'Inter Annotator Agreement (IAA) a été calculé sur 30 résumés (Fleiss

kappa : 0,72 pour les composants et de 0,68 pour la distinction preuves/conclusions) resultant dans un accord substantiel pour les deux tâches. Nous avons effectué l'annotation des relations argumentatives sur l'ensemble du corpus. L'IAA a été calculé sur 30 résumés annotés en parallèle par trois annotateurs (les deux mêmes annotateurs qui ont effectué l'annotation des composants argumentatifs, plus un annotateur supplémentaire), ce qui a donné un Fleiss kappa de 0,62. L'annotation des autres résumés a été effectuée par l'un des annotateurs susmentionnés. Nous avons proposé un pipeline complet d'extraction d'arguments pour les essais cliniques (les *preuves* et les *conclusions*), et prédisant les relations entre eux (*attaque* ou *support*). Plus précisément, notre pipeline complet d'AM pour les essais cliniques repose sur des transformateurs bidirectionnels profonds combinés à différents réseaux neuronaux, c'est-à-dire des Long Short-Term Memory (LSTM) networks, Gated Recurrent Unit (GRU) networks et Conditional Random Fields (CRFs). Nous avons procédé à une évaluation exhaustive de diverses architectures d'AM et nous avons obtenu un score F1 macro de 0,87 pour la détection des composants argumentatifs et de 0,68 pour la prédiction des relations. Notre évaluation a également révélé que les approches actuelles ne sont pas en mesure de relever adéquatement les défis posés par les textes médicaux et nous montrons que les approches basées sur les tranformers surpassent ces pipelines d'AM ainsi que les baselines standard.

2 English Abstract

Argument(ation) Mining (AM) typically aims at identifying argumentative components in text and predicting the relations among them. Evidence-based decision making (Hunter & Williams, 2012; Craven *et al.*, 2012; Longo & Hederman, 2013; Qassas *et al.*, 2015) in the healthcare domain targets at supporting clinicians in their deliberation process to establish the best course of action for the case under evaluation. However, despite its natural employment in healthcare applications, only few approaches have applied AM methods to this kind of text (Green, 2014; Mayer *et al.*, 2018, 2019), and their contribution is limited to the detection of argument components, disregarding the more complex phase of predicting the relations among them. In addition, no huge annotated dataset for AM is available for the healthcare domain. In this paper (Mayer *et al.*, 2020), we covered this gap, and we answered the following research question : *how to define a complete AM pipeline for clinical trials ?* To answer this question, we propose a deep bidirectional transformer approach combined with different neural networks to address the AM tasks of component detection and relation prediction in Randomized Controlled Trials, and we evaluate this approach on a new huge corpus of 659 abstracts from the MEDLINE database. In particular, we extended an existing dataset by annotating 500 abstracts of Randomized Controlled Trials (RCT) from the MEDLINE database, leading to a dataset of 4198 argument components and 2601 argument relations on different diseases. Following the guidelines for the annotation of argument components in RCT abstracts provided in (Trenta *et al.*, 2015), two annotators with background in computational linguistics carried out the annotation of the 500 abstracts on neoplasm. IAA among the annotators has been calculated on 30 abstracts, resulting in a Fleiss' kappa of 0.72 for argumentative components and 0.68 for the more fine-grained distinction between claims and evidence (meaning substantial agreement for both tasks). We carried out the annotation of argumentative relations over the whole dataset of RCT abstracts, including both the first version of the dataset (Trenta *et al.*, 2015) and the newly collected abstracts on neoplasm. IAA has been calculated on 30 abstracts annotated in parallel by three annotators (the same two annotators that carried out the argument component annotation, plus one additional annotator), resulting in a Fleiss' kappa of 0.62. The annotation of the remaining abstracts was carried

out by one of the above mentioned annotators. We proposed a complete argument mining pipeline for RCTs, classifying argument components as *evidence* and *claims*, and predicting the relation, i.e., *attack* or *support*, holding between those argument components. More precisely, we presented a complete AM pipeline for clinical trials relying on deep bidirectional transformers combined with different neural networks, i.e., Long Short-Term Memory (LSTM) networks, Gated Recurrent Unit (GRU) networks, and Conditional Random Fields (CRFs). We addressed an extensive evaluation of various AM architectures (e.g., for persuasive essays), and obtained a macro F1-score of .87 for component detection and .68 for relation prediction, outperforming current state-of-the-art end-to-end AM systems. Our evaluation also revealed that current approaches are unable to adequately address the challenges raised by medical text and we show that transformer-based approaches outperform these AM pipelines as well as standard baselines.

Références

CRAVEN R., TONI F., CADAR C., HADAD A. & WILLIAMS M. (2012). Efficient argumentation for medical decision-making. In *Proc. of KR 2012*, p. 598–602.

GREEN N. (2014). Argumentation for scientific claims in a biomedical research article. In *Proc. of ArgNLP 2014 workshop*.

HUNTER A. & WILLIAMS M. (2012). Aggregating evidence about the positive and negative effects of treatments. *Artificial Intelligence in Medicine*, **56**(3), 173–190. DOI : 10.1016/j.artmed.2012.09.004.

LONGO L. & HEDERMAN L. (2013). Argumentation theory for decision support in health-care : A comparison with machine learning. In *Proc. of BHI 2013*, p. 168–180.

MAYER T., CABRIO E., LIPPI M., TORRONI P. & VILLATA S. (2018). Argument mining on clinical trials. In *Proc. of COMMA 2018*, p. 137–148. DOI : 10.3233/978-1-61499-906-5-137.

MAYER T., CABRIO E. & VILLATA S. (2019). ACTA a tool for argumentative clinical trial analysis. In *Proc. of IJCAI 2019*, p. 6551–6553. DOI : 10.24963/ijcai.2019/953.

MAYER T., CABRIO E. & VILLATA S. (2020). Transformer-based argument mining for healthcare applications. In G. D. GIACOMO, A. CATALÁ, B. DILKINA, M. MILANO, S. BARRO, A. BUGARÍN & J. LANG, Éds., *ECAI 2020 - 24th European Conference on Artificial Intelligence, 29 August-8 September 2020, Santiago de Compostela, Spain, August 29 - September 8, 2020 - Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020)*, volume 325 de *Frontiers in Artificial Intelligence and Applications*, p. 2108–2115 : IOS Press. DOI : 10.3233/FAIA200334.

QASSAS M. A., FOGLI D., GIACOMIN M. & GUIDA G. (2015). Analysis of clinical discussions based on argumentation schemes. *Procedia Computer Science*, **64**, 282–289.

TRENTA A., HUNTER A. & RIEDEL S. (2015). Extraction of evidence tables from abstracts of randomized clinical trials using a maximum entropy classifier and global constraints. *CoRR*, **abs/1509.05209**.