Building Foundations for Inclusiveness through Expert-Annotated Data

Building Foundations for Inclusiveness through Expert-Annotated Data

Moreno La Quatra[1,†], Salvatore Greco[2,*,†], Luca Cagliero[2], Michela Tonti[3], Francesca Dragotto[4], Rachele Raus[5], Stefania Cavagnoli[4] and Tania Cerquitelli[2]

[1]Kore University of Enna, Enna, Italy
[2]Politecnico di Torino, Turin, Italy
[3]Università degli studi di Bergamo, Bergamo, Italy
[4]Università degli Studi di Roma Tor Vergata, Rome, Italy
[5]Università di Bologna, Bologna, Italy

Abstract

Natural Language Understanding and Generation models suffer from a limited capability of understanding the nuances of inclusive communication as they are trained on massive data, often including significant portions of non-inclusive content. Even when the models are specifically designed to address non-inclusive language detection or reformulation, they disregard, to a large extent, inclusiveness-related features that are likely correlated with the inclusive language nuances, such as the discourse type, level of inclusiveness, and intended context of use. To assess the importance of additional inclusiveness-related features, we collect a new corpus of Italian administrative documents humanly annotated by linguistic experts. Linguistic experts not only highlight non-inclusive text snippets and propose possible reformulations, but also annotate multi-aspect labels related to different inclusive language nuances. We empirically show that a multi-task learning approach that leverages the multi-aspect annotations can improve the non-inclusive text reformulation performance, thereby confirming the potential of expert-annotated data in inclusive language processing.

Keywords

inclusive language, natural language processing, text generation, deep learning

1. Introduction

Non-inclusive expressions are widespread in humanly written documents [1]. Training Natural Language Understanding and Generation models on massive data exposes them to bias issues related to language inclusiveness. Addressing this issue is particularly relevant because Artificial Intelligence (AI)-based solutions must be used responsibly to correctly model inclusive language practices and not unintentionally marginalize or disadvantage certain groups.

To mitigate the presence of bias in data, applications based on AI rely on human supervision for model training and post-processing evaluation. This is quite common in the areas of Natural Language Understanding and Generative AI, in which applications like Large Language Models (LLMs) provide end-users with conversational and language editing services [2].

The computational linguistic community has agreed on the need to leverage human expert annotations in experience-based learning for bias detection and mitigation [3, 4, 5, 6]. However, the linguistics literature often underestimates the importance of linguistic annotators because of the widespread tendency to value the figures of pre- and post-editors [7, 8]. Editing and annotation are substantially different: while language editing tools rewrite parts of the source text based on predefined expert-provided rules, Natural Language Understanding and Generation models can leverage annotations to capture the nuances of annotated text in a self-supervised manner. The use of textual annotations also relieves annotators of the task of explicitly formulating or adhering to ad hoc linguistic rules.

In the context of inclusive language understanding and generation, most of the previous work exploits rule-based or round-trip translations to annotate texts for inclusivity issues [9, 10, 11, 12]. However, these works often overlook the significance of human expert annotations, opting instead for rule-based approaches or artificially created datasets generated through round-trip translations. The role of linguistic annotators in providing specific understanding and annotations of language data is crucial for developing more inclusive AI models [13, 14].

A limited body of work has been devoted to generating and exploiting multi-faceted expert human annotations to drive AI models for inclusive language, e.g., [15, 16, 17]. However, existing benchmarks of annotated text for inclusive language processing neglect potentially relevant aspects such as the level of inclusiveness, the intended context of use, and the text genre. These aspects have the potential to improve the inclusive language understanding and generation capabilities of AI models.

This paper proposes an expert-annotated dataset covering these new aspects and investigates their usefulness in enhancing the performance of the task of non-inclusive text reformulation in the absence of rule-based editing models.

To this end, we enrich a corpus of Italian administrative documents with multi-aspect annotations, providing more insights into the inclusive language nuances. The purpose is to enable the study of new features describing inclusiveness aspects neglected by existing approaches, such as the level of inclusiveness, register, and genre. By enriching the language descriptions with new inclusiveness-related features, we provide the research community with new resources to

Published in the Proceedings of the Workshops of the EDBT/ICDT 2024 Joint Conference (March 25-28, 2024), Paestum, Italy
*Corresponding author.
†These authors contributed equally.
moreno.laquatra@unikore.it (M. La Quatra); salvatore_greco@polito.it (S. Greco); luca.cagliero@polito.it (L. Cagliero); michela.tonti@unibg.it (M. Tonti); francesca.dragotto@gmail.com (F. Dragotto); rachele.raus@unibo.it (R. Raus); stefania.cavagnoli@uniroma2.it (S. Cavagnoli); tania.cerquitelli@polito.it (T. Cerquitelli)
https://www.mlaquatra.me/ (M. La Quatra); https://grecosalvatore.github.io/ (S. Greco)
0000-0001-8838-064X (M. La Quatra); 0000-0001-7239-9602 (S. Greco); 0000-0002-7185-5247 (L. Cagliero); 0000-0001-5306-3054 (R. Raus); 0000-0002-9039-6226 (T. Cerquitelli)

enhance the understanding and writing capabilities of AI-based solutions.

We also collect preliminary results on the use of multi-aspect annotations in a multi-task learning approach to enhance non-inclusive language reformulations. The results confirm the potential of the inclusiveness-related expert annotations.

2. The annotation process

The term *annotation* is often used to indicate the process by which textual data are subjected to a tightly interrelated two-phase activity [6]: a) Identification, selection, and localisation of specific documents, and b) Interpretation and labeling of those documents. The first phase entails identifying and detailing the text segments that exhibit the linguistic phenomenon under investigation. Subsequently, in the interpretation phase, the selected occurrences are humanly labeled. These annotations may encompass various forms ranging from a selection of pre-established alternatives to free-text comments or possible reformulations.

Unlike human annotators, AI models often lack cognitive abilities such as common sense reasoning and generalization capabilities due to the relatively limited numbers of linguistic examples used for model training compared to the impressive variety of natural language forms.

Human annotators need sufficient expertise to interpret nuanced linguistic phenomena and assign appropriate labels adequately. Their annotations are at the base of a supervised learning process. The trained models can progressively learn from annotated data as automatized humans do, but at a scale not possible through manual work alone.

Annotation of Italian administrative documents. We have designed and utilized a novel benchmark dataset for inclusive language writing in Italian. This dataset comprises administrative communications sourced from the Italian public administration, spanning across both national and regional levels. We annotate the corpus at the sentence level. To this end, we set up a heterogeneous team of 13 linguistic experts with diverse experiences and expertise in inclusive language. The team consists of predominantly female individuals, all native Italian speakers. All the annotators are educated: 57% have at least 10 years of experience in linguistics, and 50% have at least 3 years of experience in inclusive language. In addition, the annotators received, on average, about 30 hours of training specific to inclusive language annotations.

Each human annotator independently assigns inclusiveness-related metadata to the document sentences. Each sentence can be enriched with multiple annotations. The annotations consist of (a) The reformulation of any non-inclusive piece of text, i.e., an alternative inclusive form; (b) The level of inclusiveness of the input sentence indicating whether a sentence is non-inclusive, inclusive, or not pertinent; (c) The register or intended context of use, i.e., *Standard*, *Specialized*, or *Informative/Educational*; (d) the discourse type or genre, i.e., *Legal*, *Administrative*, *Technical*, or *Informative/Educational*.

Additional contextual aspects could be included in future annotations to enhance models' understanding of inclusive language usage further. By jointly providing those annotations, the experts aimed to capture inclusive language's nuanced, multi-faceted nature.

By learning language inclusiveness patterns from a diversified, context-dependent set of expert annotations, AI models gain exposure to subtle interpretive differences. The consistency across annotations is ensured through detailed guidelines and instructions provided to experts. Before full annotation, a collaborative analysis of a sample set identifies any divergent interpretations to refine guidance.

Statistics on annotated data. Table 1 reports the number of annotated sentences for each aspect, separately for the training, validation, and test sets.

Task ID	Train	Validation	Test
NILR	6491	956	579
ILC	9207	1421	866
RC	2167	338	247
GC	2166	338	248

Table 1
Statistics on data. NILR=Non-Inclusive Language Reformulation, ILC=Inclusiveness Level Classification, RC=Register Classification, GC=Genre Classification.

Example of annotations. Table 2 shows an example of an Italian annotated sentence (as well as the corresponding English translation for non-Italian readers). Linguistics experts assign different annotations to each sentence. In this example, they have assigned three labels to the sentence. Regarding inclusiveness, the sentence has been categorized as non-inclusive because it contains *"Il Presidente"* (i.e., Chair/President) and *"Rettore"* (i.e., Rector), which are masculine declensions of professional roles. In addition, the sentence also contains *"suo decreto"*, which refers to a decree that comes from a male person, so the sentence is not inclusive. The discourse sequence is of the administrative type, as the content refers to an administrative topic, and the used language is specialized, as the content describes specific and technical aspects.

3. Case study: Leveraging Aspects for Italian Inclusive Language Reformulation

We conduct an empirical analysis to examine the impact of utilizing expert annotations in inclusive language generation. Specifically, we investigate the advantages of simultaneously addressing two key objectives: reformulating non-inclusive language and predicting various aspects of inclusiveness.

Tasks. Given a non-inclusive piece of text T, the *Non-Inclusive Language Reformulation* (NILR) task aims at generating an equivalent inclusive natural language form. The NILR task is a sequence-to-sequence problem, where the input is a non-inclusive sentence and the output is the corresponding inclusive sentence.

Given T and an aspect A, the goal is to predict the A's value for T. A can be the level of inclusiveness, register or intended context of use, and discourse type or genre. According to the aspect under analysis, the corresponding sub-tasks are denoted by *Inclusiveness Level Classification* (ILC), *Register Classification* (RC), and *Genre Classification*

	Sentence	Reformulation	Inclusive Class	Discursive Sequence	Clear Language
IT	*"Il Presidente, scelto dal Rettore tra i professori ordinari dell'Ateneo con competenze in ambito di valutazione, accreditamento e qualità e nominato con suo decreto, previo parere del Senato Accademico;"*	*"Chi ricopre la carica di Presidente, su scelta di chi riveste il ruolo di Rettore tra il personale docente ordinario dell'Ateneo con competenze in ambito di valutazione, accreditamento e qualità e in seguito a nomina con suo decreto, previo parere del Senato Accademico;"*	Non-inclusivo	Amministrativo	Specialistico
EN	*"The Chair/President, selected by the Rector among the full professors of the University, with expertise in the fields of evaluation, accreditation, and quality and appointed by his decree, subject to the opinion of the Academic Senate;"*	*"Who serves as Chair/President, selected by who holds the position of Rector, among the full professors of the University with expertise in the fields of evaluation, accreditation, and quality and appointed by his or her decree, subject to the opinion of the Academic Senate;"*	Non-inclusive	Administrative	Specialized

Table 2
Example of sentence annotations illustrating non-inclusive language reformulation in Italian (IT) and English (EN), along with corresponding inclusiveness classification, discursive sequence, and clear language class.

Setting	R-1	R-2	R-L	Human Eval
Single-Task	74.95	64.09	74.79	0.67
Multi-Task	**75.58**	**64.37**	**75.36**	**0.70**

Table 3
Performance comparison between Single- and Multi-task Learning approaches in inclusive language generation, evaluated based on ROUGE scores (R-1, R-2, R-L) and human evaluation.

(GC). The ILC, RC, and GC tasks are treated as separate classification problems, where the input is a sentence and the output is the corresponding aspect value.

Single- vs. Multi-Task Learning To compare the performance of models trained using different learning approaches, we conducted experiments in both single-task and multi-task learning settings.

In *Single-Task Learning*, we exclusively focus on the task of Non-Inclusive Language Reformulation (NILR), disregarding all aspect-related annotations. We leverage an encoder-decoder architecture, specifically BART-IT [18], which is a BART architecture [19] pre-trained on a clean Italian corpus [20]. The model is fine-tuned on the NILR task with the twofold objective of modifying the input sentence to make it inclusive while maintaining the original meaning.

Conversely, in *Multi-Task Learning*, we integrate the NILR task with Aspect Classification tasks during training (i.e., ILC, RC, and GC). For the additional tasks, we specifically leverage the encoder component of the model, which extracts representations of the input text. The encoder component is additionally trained with a classification objective. Each task is associated with a separate classification head, trained to predict the corresponding aspect value for the input sentence. By interleaving these tasks during training, the model learns to simultaneously address NILR and create encoder representations that capture various aspects related to inclusiveness.

Evaluation Metrics. We evaluate the quality of the text reformulation using a standard train-validation-test split on our expert-annotated data. To compare the automatically generated and expected reformulations, we use the established ROUGE F1-scores [21]. They measure the unit overlap, in terms of the number of n-grams in common, between the two pieces of text. The larger the score, the higher the syntactic similarity. R-1, R-2, and R-L count the unit overlap in terms of unigrams, bigrams, and longest common subsequences, respectively.

To complement the quantitative evaluation, we also perform a qualitative evaluation of the achieved results. We involved six human evaluators who were asked to label each model-generated sentence as: *correct* if it accurately maintained the original meaning while using inclusive language appropriately for the context; *partially correct* if some aspects were reformed correctly, but others were missed or inaccurate; or *not correct* if the rewriting fundamentally failed to capture the original meaning or usage intention. This multi-level feedback aims at capturing the models' ability to perform the rewriting task sensitively across different scenarios beyond just string-matching metrics.

To each reformulation, we assign a score to each annotation as follows: 1 for *correct*, 0.5 for *partially correct*, and 0 for *incorrect*. The final score for each reformulation is computed as the average over all the expert annotations ($m = 6$). Finally, we average the scores for all the reformulations ($n = 30$) to obtain a single score for each model.

Results' overview. Columns 2, 3, and 4 in Table 3 show the ROUGE scores for both models. The multi-task learning achieves the best performance on all the quantitative metrics. Regarding the human evaluation, we obtained 6 annotations for 30 reformulations for each model. For the model trained with the single task configuration, 93 were correct, 55 were partially correct, and 32 were incorrect. Instead, for the multi-task model, 101 were correct, 49 were partially correct, and 30 were incorrect. Column 5 reports the average human evaluation scores for both models. The human scores are coherent with the quantitative ones, showing that the model trained under multi-task settings benefits from the additional labels. Based on these preliminary results, we can conclude that the nuanced and multidimensional annotations of inclusive language have the potential to develop a more comprehensive approach to modeling inclusive language.

4. Conclusions

This paper discussed and experimentally demonstrated that the role and contribution of human annotators are of paramount importance in improving the quality of NLP results and the writing capability of generative approaches in inclusive communication. Starting from a new Italian administrative corpus, we enriched it with a variety of annotations with the help of a team of language experts. This included (i) reformulating gendered language and acronyms, (ii) rewriting to enhance readability for the visually impaired, and (iii) defining the intended context of use (register) and text genre. The preliminary experimental results on the annotated corpus are promising and highlight the potential of the newly proposed annotations to develop a more comprehensive and richer approach that improves the ability of the generative algorithm to propose comprehensive and integrative reformulations.

Limitations. i) The annotation is *language-specific*, limited to the Italian language, thereby constraining its utility in multilingual scenarios; and ii) It is *formal communication-specific*. Tailored to tackle the challenge of inclusive language in administrative and academic settings, the natural language tasks are exclusively trained on administrative documents, potentially lacking suitability for diverse contexts like legal and web communications.

Future work. As part of the E-MIMIC[1] (Empowering Multilingual Inclusive Communication) project, we are currently working on a multilingual annotation process to overcome these issues and foster inclusive communication across different domains and languages. A team of experts is annotating a large corpus of documents according to linguistic criteria to label linguistic resources in a multilingual setting.

Finally, we want to exploit text-based explainability techniques [22, 23] to perform further human validation of the models produced.

Ethical Considerations. All the gathered documents are public and therefore freely accessible on the internet. All references to proper names of people and institutions have been anonymized and replaced with random names for privacy reasons.

Acknowledgments

This study was carried out within the project "E-MIMIC: Empowering Multilingual Inclusive Communication", funded by the Ministero dell'Universitá e della Ricerca - with the PRIN 2022 (D.D. 104 - 02/02/2022) program.

References

[1] S. J. Ashwell, P. K. Baskin, S. L. Christiansen, S. A. DiBari, A. Flanagin, T. Frey, R. Jemison, M. Ricci, Three recommended inclusive language guidelines for scholarly publishing: Words matter, Learn. Publ. 36 (2023) 94–99. URL: https://doi.org/10.1002/leap.1527. doi:10.1002/LEAP.1527.

[2] A. Balayn, J. Yang, Z. Szlávik, A. Bozzon, Automatic identification of harmful, aggressive, abusive, and offensive language on the web: A survey of technical biases informed by psychology literature, ACM Trans. Soc. Comput. 4 (2021) 11:1–11:56. URL: https://doi.org/10.1145/3479158. doi:10.1145/3479158.

[3] R. Artstein, M. Poesio, Bias decreases in proportion to the number of annotators, in: Proceedings of FG-MoL 2005: The 10th conference on Formal Grammar and The 9th Meeting on, volume 139, 2009.

[4] R. Artstein, M. Poesio, Inter-coder agreement for computational linguistics, Computational linguistics 34 (2008) 555–596.

[5] J. Carletta, Assessing agreement on classification tasks: The kappa statistic, Computational Linguistics 22 (1996) 249–254. URL: https://aclanthology.org/J96-2004.

[6] P. S. Bayerl, K. I. Paul, What determines inter-coder agreement in manual annotations? a meta-analytic investigation, Computational Linguistics 37 (2011) 699–725. URL: https://aclanthology.org/J11-4004. doi:10.1162/COLI_a_00074.

[7] J. Monti, Dalla zairja alla traduzione automatica: riflessioni sulla traduzione nell'era digitale, Loffredo, 2019.

[8] P. Sánchez-Gijón, D. Kenny, Selecting and preparing texts for machine translation: Pre-editing and writing for a global audience, Machine translation for everyone: Empowering users in the age of artificial intelligence 18 (2022) 81.

[9] B. Alhafni, N. Habash, H. Bouamor, User-centric gender rewriting, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Seattle, United States, 2022, pp. 618–631. URL: https://aclanthology.org/2022.naacl-main.46. doi:10.18653/v1/2022.naacl-main.46.

[10] C. Amrhein, F. Schottmann, R. Sennrich, S. Läubli, Exploiting biased models to de-bias text: A gender-fair rewriting model, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 4486–4506. URL: https://aclanthology.org/2023.acl-long.246. doi:10.18653/v1/2023.acl-long.246.

[11] T. Sun, K. Webster, A. Shah, W. Y. Wang, M. Johnson, They, them, theirs: Rewriting with gender-neutral english, CoRR abs/2102.06788 (2021). URL: https://arxiv.org/abs/2102.06788. arXiv:2102.06788.

[12] E. Vanmassenhove, C. Emmery, D. Shterionov, NeuTral Rewriter: A rule-based and neural approach to automatic rewriting into gender neutral alternatives, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 8940–8948. URL: https://aclanthology.org/2021.emnlp-main.704. doi:10.18653/v1/2021.emnlp-main.704.

[13] A. Piergentili, D. Fucci, B. Savoldi, L. Bentivogli, M. Negri, Gender neutralization for an inclusive machine translation: from theoretical foundations to open challenges, in: Proceedings of the First Workshop on Gender-Inclusive Translation Technologies, European Association for Machine Translation, Tampere, Fin-

[1] https://dbdmg.polito.it/e-mimic/

land, 2023, pp. 71–83. URL: https://aclanthology.org/2023.gitt-1.7.

[14] M. Rosola, S. Frenda, A. T. Cignarella, M. Pellegrini, A. Marra, M. Floris, et al., Beyond obscuration and visibility: Thoughts on the different strategies of gender-fair language in italian, in: CLiC-it 2023. Proceedings of the 9th Italian Conference on Computational Linguistics. Venice, Italy, November 30-December 2, 2023., volume 3596, CEUR-WS, 2023, pp. 1–10.

[15] G. Attanasio, S. Greco, M. La Quatra, L. Cagliero, M. Tonti, T. Cerquitelli, R. Raus, E-mimic: Empowering multilingual inclusive communication, in: 2021 IEEE International Conference on Big Data (Big Data), IEEE, 2021, pp. 4227–4234.

[16] M. La Quatra, S. Greco, L. Cagliero, T. Cerquitelli, Inclusively: An ai-based assistant for inclusive writing, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2023, pp. 361–365.

[17] Raus, Rachele, Tonti, Michela, Cerquitelli, Tania, Cagliero, Luca, Attanasio, Giuseppe, La Quatra, Moreno, Greco, Salvatore, L´analyse du discours et l´intelligence artificielle pour réaliser une écriture inclusive : le projet emimic, SHS Web Conf. 138 (2022) 01007. URL: https://doi.org/10.1051/shsconf/202213801007. doi:10.1051/shsconf/202213801007.

[18] La Quatra, Cagliero, Bart-it: An efficient sequence-to-sequence model for italian text summarization, Future Internet 15 (2022) 15. URL: http://dx.doi.org/10.3390/fi15010015. doi:10.3390/fi15010015.

[19] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 7871–7880. URL: https://aclanthology.org/2020.acl-main.703. doi:10.18653/v1/2020.acl-main.703.

[20] G. Sarti, M. Nissim, It5: Large-scale text-to-text pretraining for italian language understanding and generation, arXiv preprint arXiv:2203.03759 (2022).

[21] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013.

[22] G. Sarti, N. Feldhus, L. Sickert, O. van der Wal, M. Nissim, A. Bisazza, Inseq: An interpretability toolkit for sequence generation models, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 421–435. URL: https://aclanthology.org/2023.acl-demo.40. doi:10.18653/v1/2023.acl-demo.40.

[23] F. Ventura, S. Greco, D. Apiletti, T. Cerquitelli, Trusting deep learning natural-language models via local and global explanations, Knowl. Inf. Syst. 64 (2022) 1863–1907. URL: https://doi.org/10.1007/s10115-022-01690-9. doi:10.1007/s10115-022-01690-9.