A formal and computational semantics approach to identifying underexplored topics: the case of artificial intelligence ethics charters and manifestos

Maria Boritchev[1] and Tiphaine Viard[2]
[1] S2A, IDS, Télécom Paris, Institut Polytechnique de Paris, maria.boritchev@telecom-paris.fr
[2] i[3], SES, Télécom Paris, Institut Polytechnique de Paris, tiphaine.viard@telecom-paris.fr

October 2023

In any given topic, identifying *what is absent* is a challenging task. It requires having external knowledge, and formulating hypotheses about underdiscussed topics, and then veryfying them.

In the social sciences, it is common to use *dictionary approaches*. The principle is as follows: the investigator, through their own knowledge, qualitative interviews, literature surveying and so on, identifies themes that seem underdiscussed, and drafts a list of keywords associated to each topic. The investigator then performs a search in their corpora, and validates, or not, the absence of said topics. Multiple variations and subtleties exist, though the base principle remains the same.

1 Artificial intelligence ethics charters and manifestos

Public institutions, private ones, countries, groups of countries, most of those who have to interact with artificial intelligence (AI) have in the past 10 years came up with a charter or manifesto for AI. Multiple studies provide qualitative and quantitative analyses of these guidelines, searching for common principles for AI emerging from these documents [Jobin et al., 2019, Fjeld et al., 2020, Tidjon and Khomh, 2022]. While these charters and manifestos have the same object (AI), they address it in a variety of ways, calling for more, or less regulation; for it to be considered as a tool, a threat, or a paradigmatic shift; presenting it as a universal solution, a ineluctable future, and so on.

Following the research questions raised in metanalyses of AI ethics guidelines, [Gornet and Viard, 2023] present a corpus of 74 charters and manifestos discussing AI ethics, in English. Conducting a quantitative text analysis, they show that the discussions are articulated around three mains topics: (1) social well-being, (2) company governance and (3) policy recommendations. Notably, the authors release their dataset[1], allowing for further analyses to be conducted on its contents.

Such a corpus acts as a way to capture the attitudes of the different actors that come into play regarding AI, at a given time. Analysing it provides us with a map of convergences and dissensions in the discussions surrounding AI, and helps outline the various tension points. In the context of Undone CS, we are particularly interested in developing novel methods showing the absence of discussion of some points, the lack of interest in some of the questions surrounding AI, following studies such as the one presented in [Roche et al., 2022], where the authors look for key terms in the documents to identify missing themes and show the under-representation of populations from the global south.

[1] https://mapaie.telecom-paris.fr

2 Formal and computational semantics as cartography tools

In [Gornet and Viard, 2023], the authors use textual analysis through frequency of 2 and 3-grams to identify the themes of the documents in their corpus. One of the difficulties raised by the authors has to do with the polysemy of terms such as "fairness", "beneficence", and "accountability" found in the documents. We would like to deepen the study of these documents by applying more advanced natural language processing (NLP) methods and analysing this corpus' semantic contents. To achieve this, we would like to use formal and computational semantics methods, so as to extract semantic graphs from each of the documents in the corpus, and then apply methods from graph studies to the obtained semantic graphs in order to identify the underlying structures. The formal and computational semantics tool we would use for this task is the Abstract Meaning Representation (AMR) framework [Banarescu et al., 2013]. From a linguistic perspective, AMR provides a way to model the meaning of a sentence through the use of (verbal) concepts, instances, relations, and literals. In terms of graph theory, AMR graphs are single rooted, oriented graphs, with labelled vertices and edges.

The main advantages of this approach come from the fact that AMRs are well-established in the field of deep learning semantic parsing, achieving relatively good results for sentences in English[2] with results up to 86.7% semantic match (Smatch) score [Cai and Knight, 2013] with the methods presented in [Lee et al., 2022]. The obtained representations are graphs, one per sentence or sub-sentence depending on the length of the input sentence. Once the set of graphs is obtained after parsing, formal semantics and graph theory methods can be used to study the underlying structures, similarities, and (semantic) clusters.

The limitations we envision are twofold and relate to limitations of automatic parsing. Though current deep learning AMR parsing systems do get very good Smatch scores, they are still far from perfect, and the errors tend to accumulate as the sentences get longer and more complex. This brings us to the second problem, that has to do with *coreference resolution*: as sentences and documents get longer, one has to solve coreferences (such as pronoun attributions) in order to produce the correct graphs. Systems such as the one presented in [Fu et al., 2021] provide us with a method for coreference resolution and AMR graph production at document level, while also introducing attribution errors.

The study we envision is a preliminary one, considering all the technical limitations listed above. If in this reduced setting we see the outline of interesting results, we will be able to invest time and money to perform expert linguistic annotations and improve the quality and significance of these results. In that setting, our goal is to develop AMR methods to identify more precisely missing topics, thusly expanding works on dictionary search.

About the authors

Tiphaine Viard is a *maîtresse de conférences* (roughly equivalent to tenured assistant professor) of AI and society in the social sciences laboratory at Télécom Paris/Institut Polytechnique de Paris. Her work and expertise are at the crossroads of graph modelling for real-world phenomena and expliciting the social world of artificial intelligence.

Maria Boritchev is a *maîtresse de conférences* working on NLP and computational linguistics in the computer science laboratory at Télécom Paris/Institut Polytechnique de Paris. Her work and expertise are within semantic studies of human-produced dialogue and discourse.

References

[Banarescu et al., 2013] Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., and Schneider, N. (2013). Abstract Meaning Representation for Sem-

[2] https://paperswithcode.com/task/amr-parsing

banking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics. 2

[Cai and Knight, 2013] Cai, S. and Knight, K. (2013). Smatch: an evaluation metric for semantic feature structures. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 748–752, Sofia, Bulgaria. Association for Computational Linguistics. 2

[Fjeld et al., 2020] Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., and Srikumar, M. (2020). Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for ai. *Berkman Klein Center Research Publication*, (2020-1). 1

[Fu et al., 2021] Fu, Q., Song, L., Du, W., and Zhang, Y. (2021). End-to-end amr coreference resolution. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4204–4214. 2

[Gornet and Viard, 2023] Gornet, M. and Viard, T. (2023). Mapping AI Ethics: a quantitative analysis of the plurality, and lack thereof, of discourses. In *SSRN*. 1, 2

[Jobin et al., 2019] Jobin, A., Ienca, M., and Vayena, E. (2019). The global landscape of ai ethics guidelines. *Nature machine intelligence*, 1(9):389–399. 1

[Lee et al., 2022] Lee, Y.-S., Astudillo, R., Lam, H. T., Naseem, T., Radu, F., and Roukos, S. (2022). Maximum Bayes Smatch Ensemble Distillation for AMR Parsing. In *Conference of the North American Chapter of the Association for Computational Linguistics*, pages 5379–5392, Seattle, USA. Association for Computational Linguistics. 2

[Roche et al., 2022] Roche, C., Wall, P., and Lewis, D. (2022). Ethics and diversity in artificial intelligence policies, strategies and initiatives. *AI and Ethics*, pages 1–21. 1

[Tidjon and Khomh, 2022] Tidjon, L. N. and Khomh, F. (2022). The different faces of ai ethics across the world: A principle-to-practice gap analysis. *IEEE Transactions on Artificial Intelligence*. 1