Towards a Methodology for Technoscientific Objects Extraction (Short Paper)

Alberto Cammozzo[2], Emanuele Di Buccio[1,2,3,*], Paolo Giardullo[2], Federico Neresini[2] and Andrea Sciandra[2]

[1]Department of Information Engineering, University of Padova, Italy
[2]Department of Philosophy, Sociology, Education and Applied Psychology, University of Padova, Italy
[3]Department of Statistical Sciences, University of Padova, Italy

Abstract

Media monitoring is one of the activities carried out in the research field of Public Communication of Science and Technology (PCST). This interdisciplinary research field investigates how science and technology can affect contemporary society and how society can affect science and technology. Monitoring the media discourse can be beneficial to understanding the narrative that might affect society's perception of an issue when carried on by non-experts. One of the necessary tasks when following the discussion on the media is the automatic extraction of the actors involved. Besides people, companies, or institutions, a crucial task is the extraction of other non-human actors that play a leading role in the science narrative, such as relevant scientific terms or technologies. This paper documents our ongoing effort in extracting those terms and how they can be helpful for researchers in PCST.

Keywords

Terminology Extraction, Expert Users, Digital Sociology

1. Introduction

Science and technology have an important role in society, as shown, for instance, by the recent COVID-19 pandemic or by the increasing attention to controversial issues related to Artificial Intelligence. Indeed, science and technology, hereafter denoted as "technoscience" [1], can affect society or be affected by society, e.g., because of the public discourse on technoscientific issues. Public Communication of Science and Technology (PCST) is an interdisciplinary research field that studies the relationship between science and society and includes several activities, such as Media Monitoring, that can help to understand and follow the narrative on some issues and how that narrative might affect the public perception. The research activities performed by PCST scholars can involve studying actors involved in public debate and the relationship among those actors. When working on the Media, automatic approaches can be adopted to follow the

3rd International Conference on "Multilingual digital terminology today. Design, representation formats and management systems" (MDTT) 2024, June 27-28, 2024, Granada, Spain.
*Corresponding author.
alberto.cammozzo@unipd.it (A. Cammozzo); emanuele.dibuccio@unipd.it (E. Di Buccio); paolo.giardullo@unipd.it (P. Giardullo); federico.neresini@unipd.it (F. Neresini); andrea.sciandra@unipd.it (A. Sciandra)
0000-0003-1551-0022 (A. Cammozzo); 0000-0002-6506-617X (E. Di Buccio); 0000-0002-2560-9088 (P. Giardullo); 0000-0003-3918-2588 (F. Neresini); 0000-0001-5621-5463 (A. Sciandra)

discourse on more traditional ones, such as newspapers, or recent ones, like Social Networks. Once actors have been "extracted," they can be used as terms to formulate queries for accessing documents relevant to some specific technoscientific issues or to analyze the prominence of the actors over time or the context where they occur. Therefore, PCST scholars need methods to extract actors from possibly vast amounts of data that are infeasible to process manually.

Automatic extraction of terms from text [2] and Named Entity Recognition (NER) are therefore helpful to support PCST scholars tasks. However, NER methods are usually adopted to extract entities such as people (names), locations, and time. There are scenarios where those approaches must be tailored to extract entities to specific domains. This is the case, for instance, of the medical domain [3] or the technology domain [4] in patent corpora. In this work we are interested in non-human actors, specifically those playing a leading role in the technoscience narrative, such as relevant scientific terms or technologies.

The work reported in [4] is particularly relevant to ours since it is focused on technology extraction. In that paper, the authors investigated diverse Natural Language Processing (NLP) approaches for extracting technologies from patent data. They categorized the three approaches as gazetteer-based, rule-based, and distributional-based. As for the first approach, gazetteer-based, two different sources were used: Wikipedia (the list of emerging technologies) and O*NET, a free online text database containing job definitions and technologies related to the occupational domain. The adopted rule-based approaches included one based on lexico-syntactic patterns for hyponymy extract from data [5]. As for the third approach, they used distributional-based approaches relying on BERT for the embedding, a 2-layer bi-directional LSTM on top of the embeddings and conditional random fields, with an architecture similar to that used in [6]. Models were trained by data gathered from the proposed rule-based approaches.

In this work, we are interested in terms related to technoscience, including but not limited to technologies. We will propose a methodology for single-word and MWE extraction that, as in [4], relies on multiple approaches and multiple sources of evidence. The methodology is intended both as a first approach to extracting technoscientific objects and to support the creation of a labeled set that will be used in future work, e.g., for fine-tuning models. This paper describes the methodology and some preliminary results.

2. Methodology

This work will consider online newspaper articles as a source for terms. Even if we are working on methodology instantiations for diverse languages, such as English and Italian, this section will refer to a corpus of 260627 articles published in eight Italian newspapers in 2022. Those articles have been gathered thorugh a Media Monitoring platform called TIPS (Technoscienfic Issues in the Public Sphere).[1] The platform, originally described in [7], has been extended with a number of additional modules and services [8, 9]. In [8], a description of *hactar*,[2] the platform for collecting, extracting, cleaning, and processing online web pages from newspapers, blogs, and websites, was provided. *Hactar*, released under AGPL, is equipped with functionalities for NER thorugh the adoption of open source libraries. The corpus considered in this paper

[1] http://www.tipsproject.eu/tips
[2] https://gitlab.com/mmzz/hactar

is constituted of articles from the following newspapers: “Avvenire”, “Corriere della Sera”, “il Giornale”, “Il Mattino”, “Il Messaggero”, “la Repubblica”, “Il Sole 24 Ore”, and “La Stampa”. The URLs of the articles have been made available through Zenodo [10].

Our methodology for technoscientific object extraction focuses on two types of terms: single-word and multi-word expressions. For both the term types, we will use multiple strategies as done in [4], which includes Gazetteers, rule-based named entity candidates identification, and Supervised Machine Learning (ML) techniques. Since most of the available models have been trained for the well-known classes of entities, such as Person, Organization, and Place, our approach have two goals:

- obtaining a set of technoscientific objects without fine-tuned ML-based NER;
- building a labeled set that can be later adopted to train and evaluate fine-tuned ML-based NER approaches to extract technoscientific objects directly.

Scientific instruments, laboratory equipment, measuring instruments, medical devices, pharmaceuticals, methodologies, and disciplines are examples of “objects” of interest.

Our current approach involves common methodology steps for both types of terms:

1. **Corpus cleaning**, which involved the removal of documents in other languages, duplicates, and near-duplicates; the last step was performed using the min-wise independent permutations locality sensitive hashing scheme (MinHash) [11], relying on the implementation available in the datasketch library.[3]
2. **Extraction of candidate terms** by Part of Speech (PoS) Taggers (for single-words), also in conjunction with an approach for Multiword expression (MWE) extraction.
3. **Filter Hapax Legomena and known terms** from other “classes” using gazetteers or lists of terms obtained via ML-based NER approaches, Web API (e.g., Scopus), or looking at DBPedia entries categories.
4. **Technoscientific content classification** via Supervised Machine Learning (ML) techniques.
5. **Candidate terms ranking** by the approach proposed in [12], which requires information on terms (statistics) on documents relevant and non-relevant to technoscience.
6. **Extraction of the final term list** by manual inspection or by adopting a threshold based on term statistics or the score from the previous step.

In this paper, we will focus on steps 2–5.

2.1. Candidate Term Extraction and Filtering

As for the extraction of candidate terms (step 2), we adopted two approaches: one to extract single-word and one to extract MWEs.

As for single words, we relied on PoS Taggers to extract all the nouns occurring in the documents. The extraction of nouns was performed by Tint [13], which is based on the Standard CoreNLP Library [14] and it was explicitly developed for Italian. We considered all the terms labeled by the PoS tagger with the tags S (common noun) and SP (proper noun).

[3] https://github.com/ekzhu/datasketch

As for MWEs, we relied on NPFST [15] and its implementation available in the `phrasemachine` library.[4] The required input is, for each sentence, the set of constituting tokens and the corresponding PoS tags. To extract PoS tags, we relied on the spaCy Library because it is the tool currently adopted in the TIPS pipeline to process all the collected articles. The output of this step is a set of candidate MWEs, where some may overlap or be nested – e.g., one is a sub-string of the other – or might be closely related. Examples in Italian are:

- "telescopio spaziale James",
- "telescopio spaziale James Webb",
- "super telescopio spaziale James",
- "super telescopio spaziale James Webb".

These MWEs can be combined into a single MWE "telescopio spaziale James Webb" ("James Webb Space Telescope").

After extracting single-word and multi-word technoscientific object candidates, we filter out Hapax Legomena and terms present in available gazetteers or previously manually labeled sets, thus restricting the number of nouns and MWE to process. The resources adopted were:

- a "standard" Italian stoplist;[5]
- a list of all countries and over eleven million placenames made available in the GeoNames geographical database;[6]
- OpenStreetMap[7] nodes, relations and ways (in Italy);
- a list of persons, organizations, and places identified by the spaCy NER for Italian;
- a list of person surnames (used to filter single-word entries);
- a list of persons which includes scientists extracted through the Scopus Web API, which were manually checked for a previous study [16];
- a list of journals gathered from Scopus;
- a list of farmaceutical corporations, a list of drugs and drug active ingredients gathered from the Website of the Agenzia Italiana del Farmaco[8].

Note that the drugs and active ingredients mentioned in the last item are objects of interest that can be added to the final list. When processing single-words and MWEs, we also checked for corresponding entries in DBPedia through the DBPedia Lookup service[9], looking for the presence of the term in the label of the returned entities; this approach helped to identify some persons not present in our gazetteers and the category of some technoscientic terms, e.g., proteins or chemical compounds.

After filtering, the remaining MWEs required a subsequent step because of related terms. Following the suggestion reported in [15] on merging related terms, we added a subsequent step for grouping terms. In that work, a high-level algorithm is illustrated; we opted for a

[4] https://github.com/slanglab/phrasemachine
[5] https://github.com/stopwords-iso/stopwords-it
[6] https://www.geonames.org
[7] https://www.openstreetmap.org/
[8] https://www.aifa.gov.it/liste-dei-farmaci
[9] https://www.dbpedia.org/resources/lookup/

different approach that used both LSH with MinHash and the Morrone's Index [17]. LSH with MinHash was adopted to group related MWEs. We adopted the implementation available in the datasketch library, using 128 permutations and MWE representations based on 4-characters long n-grams, with a threshold of 0.5. The nearest MWEs by MinHash were then compared using the Morrone's Index. The index provides a measure of the extent to which a MWE is significant in a corpus, and it is defined as follows:

$$IS = P * \sum_{i=1}^{L} \frac{f_{\text{MWE}}}{f_{t_i}} \tag{1}$$

where L is the number of tokens in the MWE, f_{MWE} is the frequency of the MWE in the corpus computed as the number of documents where it occurs (document frequency), f_{t_i} is the (document) frequency of the token t_i, and P is the number of non-stopwords in the MWE. We used the standardized version of the index obtained by dividing Eq. 1 by L^2. For instance, for the previously mentioned MWEs, the value of the index are:

- "telescopio spaziale James": 0.039
- "telescopio spaziale James Webb": 0.069
- "super telescopio spaziale James": 0.0033
- "super telescopio spaziale James Webb": 0.006

Therefore, among the MWEs in this group, we will opt for the one with the highest score, i.e., "telescopio spaziale James Webb". Note that "James Webb" will be kept as an instance of a Person entity and "telescopio" (telescope) as an instance of a single-word technoscientific object.

2.2. Candidate Terms Ranking

As for step 4, since our goal is to follow the technoscientific discourse, we need a definition of the object of analysis. The work reported in [8] describes and uses a "pragmatic" solution that assumes the point of view of a hypothetical "typical newspaper reader" and what this person might recognize as "technoscientific". This solution suggests several features, including the occurrence of scientists/engineers, a discovery, a scientific instrument, or a general reference to research processes and technological innovations; these features were used to define the criteria for manually labeling documents according to their relevance to technoscience. The initial labeled set described in [8] was then extended by labeling additional documents; the final labeled set was then used to train a classifier using supervised ML. The most effective technique among those studied in [9] is based on the stacking [18] of a Multinomial Naive Bayes classifier and Logistic Regression with Coordinate Descent Methods [19]. This result was consistent for both Italian and English. Therefore, this is the approach adopted in this work to determine articles relevant and not relevant to technoscience.

Once all the articles have been classified, we ranked all the terms by the following score [12]:

$$a_t = w_t(p_t - q_t) = \log \frac{p_t(1-q_t)}{q_t(1-p_t)} (p_t - q_t) \tag{2}$$

where p_t (q_t) is the probability that a given relevant (nonrelevant) document is assigned the term t. For instance, the top 10 single words extracted from articles published in 2022 and ranked by this score are:

- “biomarcatore” (biomarker)
- “citochina” (Cytokine)
- ‘neurology”
- “her2”
- “esmo”
- “nivolumab”
- “interferone” (interferon)
- “statine” (Statins)
- “monoterapia”
- “trastuzumab”

As a proof of concept, we examined the top 1000 single words extracted, and 32 of them were not correct; errors included surnames of scientists, acronyms of scientific journals such as PNAS or PLOS, Twitter accounts (of scientists or personalities related to technoscience), and the abbreviation of organizations; some of those terms that will be added to the gazetteers. We are currently evaluating a large sample of the extracted terms.

3. Final remarks and Future Works

In this paper, we proposed a methodology for technoscientific object extraction relying on multiple strategies, which include the use of gazetteers and ML-based NER to filter out non-relevant objects from those identified by PoS taggers or by Multiword expression (MWE) extraction approaches. We provided a method for ranking the remaining terms that can be used to carry out the analysis by social science researchers, such as PCST scholars, and to build a labeled set to investigate methods for a fully automated process.

To provide access to the gazetteers and the extracted terms both for the extraction procedure and support analysis, we used *elasticsearch*,[10] thus allowing terms and MWEs (and their statistics) to be retrieved using full-text and fuzzy search both on the constituting tokens and the PoS tags. In addition, we stored the occurrence of these terms in the newspaper articles, so that we could search for technoscientific objects and monitor their presence and their relationship over time. Moreover, the study of the occurrence of technoscientific objects in conjunction with indicators such as the one on *risk* [7], might help us to gain some insights on the perception of some of these objects or how they are discussed in the media sphere.

Our current effort is devoted to a more extensive evaluation using expert annotators, i.e., researchers in PCST and Science and Technology Studies. Moreover, we will complement the methodology with ML-based approaches to NER, such as those devised to extract entities from specific domains.

[10] https://www.elastic.co/elasticsearch

References

[1] B. Latour, Science in action: How to follow scientists and engineers through society, Harvard university press, 1987.

[2] G. M. Di Nunzio, S. Marchesin, G. Silvello, A systematic review of automatic term extraction: What happened in 2022?, Digital Scholarship in the Humanities 38 (2023) I41–I47. doi:10.1093/llc/fqad030.

[3] G. M. Di Nunzio, S. Marchesin, G. Silvello, Terminology extraction in electronic health records. the examode project (poster), in: G. M. D. Nunzio, G. M. Henrot, M. T. Musacchio, F. Vezzani (Eds.), Proceedings of the 1st International Conference on Multilingual Digital Terminology Today, Padua, Italy, June 16-17, 2022 (hybrid event), volume 3161 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022. URL: http://ceur-ws.org/Vol-3161/poster1.pdf.

[4] G. Puccetti, V. Giordano, I. Spada, F. Chiarello, G. Fantoni, Technology identification from patent texts: A novel named entity recognition method, Technological Forecasting and Social Change 186 (2023). doi:10.1016/j.techfore.2022.122160.

[5] M. A. Hearst, Automatic acquisition of hyponyms from large text corpora, volume 2, Association for Computational Linguistics, 1992, p. 539. URL: http://portal.acm.org/citation.cfm?doid=992133.992154. doi:10.3115/992133.992154.

[6] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, Association for Computational Linguistics, 2018, pp. 2227–2237. URL: http://aclweb.org/anthology/N18-1202. doi:10.18653/v1/N18-1202.

[7] E. Di Buccio, A. Lorenzet, M. Melucci, F. Neresini, Unveiling latent states behind social indicators, in: R. Gavaldà, I. Zliobaite, J. Gama (Eds.), Proceedings of the First Workshop on Data Science for Social Good co-located with European Conference on Machine Learning and Principles and Practice of Knowledge Dicovery in Databases, SoGood@ECML-PKDD 2016, Riva del Garda, Italy, September 19, 2016, volume 1831 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2016. URL: https://ceur-ws.org/Vol-1831/paper_6.pdf.

[8] A. Cammozzo, E. Di Buccio, F. Neresini, Monitoring technoscientific issues in the news, in: ECML PKDD 2020 Workshops - Workshops of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2020): SoGood 2020, PDFL 2020, MLCS 2020, NFMCP 2020, DINA 2020, EDML 2020, XKDD 2020 and INRA 2020, Ghent, Belgium, September 14-18, 2020, Proceedings, volume 1323 of *Communications in Computer and Information Science*, Springer, 2020, pp. 536–553. doi:10.1007/978-3-030-65965-3_37.

[9] E. Di Buccio, A. Cammozzo, F. Neresini, A. Zanatta, TIPS: search and analytics for social science research, in: L. Tamine, E. Amigó, J. Mothe (Eds.), Proceedings of the 2nd Joint Conference of the Information Retrieval Communities in Europe (CIRCLE 2022), Samatan, Gers, France, July 4-7, 2022, volume 3178 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022. URL: https://ceur-ws.org/Vol-3178/CIRCLE_2022_paper_33.pdf.

[10] A. Cammozzo, E. Di Buccio, P. Giardullo, F. Neresini, A. Sciandra, Data from: Towards a Methodology for Technoscientific Objects Extraction (Short Paper), 2024. doi:10.5281/zenodo.10869937.

[11] A. Z. Broder, M. Charikar, A. M. Frieze, M. Mitzenmacher, Min-wise independent permutations, Journal of Computer and System Sciences 60 (2000) 630–659. doi:10.1006/jcss.

1999.1690.
[12] S. Robertson, On term selection for query expansion, Journal of Documentation 46 (1990) 359–364. URL: https://www.emerald.com/insight/content/doi/10.1108/eb026866/full/html. doi:10.1108/eb026866.
[13] A. Palmero Aprosio, G. Moretti, Tint 2.0: an All-inclusive Suite for NLP in Italian 10 (2018) 12.
[14] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, D. McClosky, The Stanford CoreNLP natural language processing toolkit, in: Association for Computational Linguistics (ACL) System Demonstrations, 2014, pp. 55–60. URL: http://www.aclweb.org/anthology/P/P14/P14-5010.
[15] A. Handler, M. Denny, H. Wallach, B. O'Connor, Bag of what? simple noun phrase extraction for text analysis, Association for Computational Linguistics, 2016, pp. 114–124. URL: http://aclweb.org/anthology/W16-5615. doi:10.18653/v1/W16-5615.
[16] F. Neresini, P. Giardullo, E. Di Buccio, B. Morsello, A. Cammozzo, A. Sciandra, M. Boscolo, When scientific experts come to be media stars: An evolutionary model tested by analysing coronavirus media coverage across italian newspapers, PLoS ONE 18 (2023). doi:10.1371/journal.pone.0284841, all Open Access, Gold Open Access, Green Open Access.
[17] A. Morrone, Temi generali e temi specifici dei programmi di governo attraverso le sequenze di discorso, in: L'attività dei governi della Repubblica italiana (1948–1994), Il Mulino; Bologna, 1996, p. 351–369.
[18] D. H. Wolpert, Stacked generalization, Neural Networks 5 (1992) 241–259. doi:10.1016/S0893-6080(05)80023-1.
[19] H.-F. Yu, F.-L. Huang, C.-J. Lin, Dual coordinate descent methods for logistic regression and maximum entropy models, Machine Learning 85 (2011) 41–75. URL: https://doi.org/10.1007/s10994-010-5221-8. doi:10.1007/s10994-010-5221-8.