

# Key Environmental Lexicon Extraction Using Generative Transformer (Short Paper)

Tomara Gotkova<sup>1</sup>, Alexander Shvets<sup>2</sup>

<sup>1</sup>Université de Lorraine, CNRS, ATILF (Nancy, France)

<sup>2</sup>Pompeu Fabra University, NLP Group (Barcelona, Spain)

## Abstract

This paper presents a study of the core environmental lexicon at the intersection of the fields of terminology and natural language processing. The goal was to find a way of automatizing the expansion of the preselected keyword list, and, in particular, to evaluate the ability of generative transformers to extract keywords unseen during the training phase. As a starting point, we collected keywords pertinent to the environmental discourse. Additionally, we compiled a corpus of texts on current and emerging environmental issues. These materials were used to train deep generative models of two types: a T5 transformer and a pointer-generator network pretrained for concept extraction as a baseline. We show that T5 significantly outperforms the baseline in detecting unseen keywords. We further provide qualitative analysis of the outcome of the resulting model applied to weakly annotated texts and confirm that the model helps to discover more keywords pertinent to the environmental topic.

## Keywords

environmental terminology, deep generative models, keyword extraction, specialized corpus

## 1. Introduction

Our primary objective is rooted in terminology: we aim to identify the **core** environmental terminology which we see as a set of central terms that shape the modern environmental discourse. As a first step towards this objective, we opt for a supervised machine learning approach that consists in training deep generative models with preselected lexical material and a specialized corpus of environmental texts. In the following sections, we comment on the theoretical framework that underlies our terminological tasks, describe the dataset, the selection of generative models, preliminary extraction results and points for future work.

## 2. Theoretical framework

### 2.1. The notion of “environmental coreness”

Environmental terminology is a patchwork of terms which belong to different disciplines (anthropology, chemistry, biology, ecology, physics) and topics (renewable energy, ocean pollution,

---


*2nd International Conference on "Multilingual digital terminology today. Design, representation formats and management systems" (MDTT) 2023, June 29–30, 2023, Lisbon, Portugal*

✉ tomara.gotkova@univ-lorraine.fr (T. Gotkova); alexander.shvets@upf.edu (A. Shvets)

ORCID 0000-0002-9098-5725 (T. Gotkova); 0000-0002-8370-2109 (A. Shvets)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

biodiversity). Due to such heterogeneity, environmental terminology defies clear-cut segmentation when it comes to certain tasks. While it is relatively easy to discern terms specific to a given environmental subtopic or subdiscipline, detecting terms which are relevant for most of the subtopics or subdisciplines at once remains a challenging (but feasible) task. For instance, [1] proposes a method of identifying *general environmental lexicon* which “cuts across the entire field of the environment”, e.g., *biologist*, *ecosystem*, *green*.

Previous research explored the notion of “coreness” as applied to both general and specialized lexicon. Depending on the purpose of a given core wordlist, core words can be defined by such properties as frequency, commonness, universality, semantic primitiveness, etc. [2, 3, 4]. We focus on the notion of environmental coreness in the specialized texts on the *current and emerging environmental issues*, e.g., air pollution, loss of biodiversity, waste management, etc. A given environmental term is considered core if it meets the following criteria: (i) it refers to the most essential environmental concept (*sustainable*), (ii) it is pertinent to several environmental subtopics at once (*ecosystem*), (iii) it exhibits strong semantic connections with other environment-related terms, (iv) it is not specific to specialized environmental discourse only as it is diffused in general language discourse as well (mass media texts, general public communication, etc.).

## 2.2. Term vs. keyword

We advocate for the lexico-semantic approach to terminology which treats terms as lexical units [5]. According to the integral component of the Meaning-Text theory – the Explanatory Combinatorial Lexicology – a lexical unit is a word which corresponds to one specific sense [6]. Hence, the term *carbon* implies a pointer to a specific sense, e.g., ‘chemical element C’. As regards to machine learning tasks, however, we deliberately refrain from using the notion of “term” to demonstrate that we stay at the level of abstract units with no clear terminological status. Instead, we use the notion of *keyword* which is a wordform devoid of clear semantic features as there is no direct reference to a specific sense. For instance, the keyword *carbon*<sup>1</sup> per se does not refer to any specific sense but it can acquire semantic features in context. It should be noted that our notion of keyword is different from the uses which refer to concepts rather than semantically ambiguous wordforms, e.g., keywords of a scientific paper.

## 3. Methodology

### 3.1. Data

**Keyword list.** As a result of continuous sampling of environmental lexical material<sup>2</sup>, we compiled the initial list of 268 unique environment-related keywords. These keywords were further divided into two categories. We selected 104 keywords which we see as core-candidates, i.e., keywords which may potentially be validated as core environmental terms (*carbon*, *climate*, *global warming*, *greenhouse gas*). The supervised models are expected to expand this list.

---

<sup>1</sup>Terms are written in italics; keywords are written in teletype.

<sup>2</sup>The process included both manual and automatic selection and was partly done in collaboration with an expert in green chemistry and an expert in lexicology [7].

The remaining 164 keywords were categorized as supplementary<sup>3</sup>. We considered the following keywords as supplementary: complex keywords built with Ckeywords (`air pollution`, `anthropogenic carbon dioxide`, `atmospheric warming`) and keywords which would not satisfy the criteria of coreness but are nevertheless important for environmental discourse (`ice`, `Earth`).

**Specialized corpus.** Our specialized corpus is a monolingual English domain-specific corpus composed of 44 reports issued by international environmental organizations such as European Environmental Agency, Intergovernmental Panel on Climate Change, United Nations Environmental Program and World Meteorological Organization. These reports give a comprehensive overview of the current and emerging environmental issues.

We converted documents to plain text excluding figures and tables and manually cleaned artifacts remained after the conversion. Consider an example of a sentence with Ckeywords given in bold and Skeywords underlined:

*Moreover, the degradation of wetlands releases **stored carbon**, fuelling **climate change**.*

### 3.2. Automatic and gold-standard annotation

We designed a simple procedure to annotate the entire corpus of about 30K sentences to have enough data samples for training a neural network. The first step is to parse the corpus using UDPipe<sup>4</sup>, while the second is to consider all the sequences of tokens of lengths from one to six (i.e., up to the maximum number of words in keywords in our lists) taking the normal forms of lexical items using their lemmas, and looking them up (with conditions on part-of-speech tags) in the lists of keywords which we automatically expanded with alternatives beforehand (e.g., for `biodiversity conservation` we added `conservation of the biodiversity`, for `bio-based - biobased`, etc.). Finally, each sentence with the corresponding found items made a single data sample. The obtained samples cover 103 out of 104 Ckeywords (`carbon-free` was not found in this corpus) and, in total, 255 out of 268 keywords. The search procedure took into account many possible occurrences including the cases of overlapping and discontinuous keywords such as `soil pollution` and `air pollution` in a phrase “`soil and air pollution`”.

Resulting samples were shuffled and split into the training, development (dev), and test subsets in the proportion 80/10/10. We performed shuffling several times until the examples were distributed among the subsets in such a way that only 80% of the keywords are used for training (they also appear in two other subsets), while other 10% and 10% are used exclusively in the dev and test subsets without intersections<sup>5</sup>. We preserve these 20% of keywords to assess the ability of the model to extract “new” keywords unseen during the training. We leverage the dev set to select the most prominent intermediate states of the model obtained during the training, and the test set – for the final evaluation. Sentences without keywords were also added proportionally to the subsets to guide the model when it should not extract anything.

---

<sup>3</sup>Further in the text, core-candidate keywords and supplementary keywords are called *Ckeyword* and *Skeyword* respectively.

<sup>4</sup><https://ufal.mff.cuni.cz/udpipe>

<sup>5</sup>A couple of thousands of samples were removed from the dataset to meet the condition of exclusiveness.

In addition to our simple automatic annotation, we manually selected and examined 200 sentences from the corpus (excluded from the subsets) and created fully annotated samples (with some keywords beyond the existing lists) that we refer to as a gold standard. The size of the overall dataset is shown in Table 1.

	Ckey+Skey	Ckey	Ckey new	# pos	# neg
Training	24,565 (206)	17,618 (80)	-	10,301	8,803
Dev	3,711 (184)	2,723 (83)	183 (9)	1,449	1,149
Test	3,703 (172)	2,737 (80)	183 (9)	1,505	1,117
Gold	592 (238)	192 (37)	35 (9)	100	100

**Table 1**

Statistics over the dataset. Ckey+Skey/Ckey/Ckey new – number of keyword occurrences in a subset (number of unique keywords in parenthesis); # pos, # neg – number of samples w/ and w/o keywords

### 3.3. Generative extraction models

The overlapping and discontinuous keywords in environmental texts create a problem in applying traditional sequence labelling-based extractors. Instead, in this work, we opt for deep neural generative models that are capable of translating a sentence into an arbitrary sequence of words (not necessarily coherently connected) like T5 [8] as we would like a model to output keywords in the form appeared in a sentence one after another separated by a reserved symbol<sup>6</sup>.

In our experiments, we worked with two versions of the pretrained transformer T5, *T5-small* and *T5-large*<sup>7</sup>. We also chose a pretrained pointer-generator-based concept extraction model (CE-PGN) [10] as an alternative that we successfully applied for public discourse analysis in the domain of interior and urban design [11]. Originally, this model was designed to extract concepts mainly in a form of nominal phrases which is not the exclusive form for the keywords considered in this work. Still, we assumed that tuning it on our data could change its behaviour.

## 4. Results

We report on the precision  $P = TP/(TP + FP)$  and recall  $R = TP/N_P$  scores for different types of keywords (Skeywords, Ckeywords, and *Ckeywords new* – Ckeywords unseen during the training) in Table 2 where  $TP$  is the number of correctly extracted mentions of the scored type,  $FP$  – the number of extracted mentions out of all ground-truth mentions ( $FP$  does not depend on the type under scoring), and  $N_P$  – the number of ground-truth mentions of the scored type.

As expected, the original CE-PGN model extracts a small number of keywords with a very low precision as it tends to find all the concepts independently of the domain. The fine-tuning

<sup>6</sup>E.g., Sustainable forest management can maintain... → Sustainable \* Sustainable forest management \* forest

<sup>7</sup>For languages other than English, mT5 shall be used as it allows for cross-lingual transfer learning [9].

<sup>8</sup>The model was tuned on the same training set but annotated only with Ckeywords.

	Dev						Test						Gold					
	Ckey+Skey		Ckey		Ckey new		Ckey+Skey		Ckey		Ckey new		Ckey+Skey		Ckey		Ckey new	
	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R
CE-PGN	5.96	31.66	4.27	30.41	0.36	36.61	6.06	32.40	4.55	32.37	0.50	50.82	25.73	36.99	9.71	35.42	2.62	48.57
CE-PGN tuned	94.94	87.47	93.18	86.82	0.57	0.54	95.11	85.53	93.41	84.40	2.40	2.19	81.23	51.18	67.29	75.00	4.11	8.57
T5-small tuned	84.64	85.85	80.73	88.91	12.95	46.99	89.60	83.53	86.89	86.92	22.46	<b>56.83</b>	82.45	52.36	71.79	87.50	20.48	48.57
T5-large tuned	88.13	<b>91.05</b>	84.90	<b>93.94</b>	22.09	<b>70.49</b>	94.50	<b>92.28</b>	92.80	<b>93.68</b>	32.08	51.36	82.82	<b>59.46</b>	70.68	<b>91.67</b>	24.74	<b>68.57</b>
T5-large c-tuned <sup>8</sup>	93.45	69.23	93.41	93.65	29.96	42.08	93.02	67.65	93.02	91.49	26.27	36.61	85.31	46.11	78.64	90.10	32.86	65.71

**Table 2**  
Results breakdown

re-oriented it towards the environmental domain – both scores were significantly improved. However, it performed poorly on extracting unseen keywords.

T5-large performed better than other models apart from when the small version gained a slightly higher recall score on the unseen Ckeywords of the test set. Interestingly, the annotation with Skeywords helped to detect Ckeywords better. The model that was trained only to extract Ckeywords (T5-large c-tuned) generalized poorer and missed many more unseen Ckeywords.

For the quality check of the extraction results, we manually checked 171 non-annotated keywords extracted from dev set using T5-large. As a result, 70 novel keywords (41%) were obtained, other 32 (19%) corresponded to existing keywords missing in automatic annotation due to mistakes of the parser, and only the rest 69 (40%) were false negatives, i.e., not keywords. 45 keywords out of the 70 novel were combinations of already existing keywords in our lists (ecological drought, biomass contaminant), the remaining 25 keywords were new to us (smog, renewable electricity, biomethane). This result is linguistically valuable for us: all 25 new keywords are pertinent to the environmental topic. Although, some keywords are too specific (cryosphere), all of them are considered as an important addition to our list.

## 5. Conclusions and Future Work

Results of our experiments provided several valuable insights as regards both linguistics and information extraction areas. First, the preselected keywords proved pertinent to the environmental topic and, in particular, to the vocabulary of the current and emerging environmental issues. More specifically, only 13 keywords out of 268 were not present in our corpus (5%). Second, tests performed with T5-large demonstrated that supplementary lexical material (Skeywords) enhanced the model’s ability to detect Ckeywords. Therefore, as the list of Ckeywords used to train the model grows, it is necessary to add to the list of Skeywords as well. Third, we consider it now important to increase the number of manually annotated samples to improve the gold standard dataset and this will allow us to train the model on annotated data of high quality in addition to automatically annotated sets. Fourth, T5-large model proved efficient for extracting unseen keywords: it detected 50-70% of them in a set (62% across all the evaluation sets). Finally, we extracted 70 novel keywords pertinent to the topic of current and emerging environmental issues which were not present in our preselected keyword list.

The ultimate goal of our research, which goes beyond this study, is multifold. The finalized

keyword list will be used to scrape data from social networks, namely Twitter and Reddit, to monitor the general public's perception of core environmental terms. As a parallel task, both preselected and extracted keywords will be subject to a lexicographic analysis in order to convert them into meaningful lexical units and describe them in a lexicographic resource. In some cases, a given complex keyword may be decomposed into several terms. For example, the keyword `climate pollutant` should be converted to and lexicographically described as two separate terms *climate* and *pollutant*, for phraseological reasons. Additionally, the obtained list of terms will be analyzed according to the criteria of environmental coreness discussed in 2.1. If a given term satisfies all four criteria, it can be validated as a core environmental term. We would also like to explore the differences in terminology in multilingual material with mT5 and study the transferability of the obtained extractive models to other domains.

## Acknowledgements

This research was funded by the EC-funded research and innovation programme Horizon Europe under the grant agreement number 101070278 and by the French PIA project "Lorraine Université d'Excellence", reference ANR-15-IDEX-04-LUE.

## References

- [1] P. Drouin, M.-C. L'Homme, B. Robichaud, Lexical profiling of environmental corpora, in: N. C. C. chair), K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odiijk, S. Piperidis, T. Tokunaga (Eds.), Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), European Language Resources Association (ELRA), Paris, France, 2018, pp. 3419–3425.
- [2] R. Carter, Is there a Core Vocabulary? Some Implications for Language Teaching\*, Applied Linguistics 8 (1987) 178–193.
- [3] E. Zenner, D. Speelman, D. Geeraerts, Core vocabulary, borrowability and entrenchment: A usage-based onomasiological approach, Diachronica 31 (2014) 74–105.
- [4] V. Brezina, D. Gablasova, Is There a Core General Vocabulary? Introducing the New General Service List, Applied Linguistics 36 (2013) 1–22.
- [5] M.-C. L'Homme, Lexical semantics for terminology : an introduction, Terminology and lexicography research and practice (TLRP), John Benjamins Publishing Company, Amsterdam Philadelphia, 2020.
- [6] I. A. Mel'čuk, A. P. Clas, A. Polguère, Introduction à la lexicologie explicative et combinatoire, Universités francophones, Duculot, 1995.
- [7] T. Gotkova, N. Chepurnykh, Public perception and usage of the term : Linguistic analysis in an environmental social media corpus, Psychology of Language and Communication 26 (2022) 297–312.
- [8] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, Journal of Machine Learning Research 21 (2020) 1–67.
- [9] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, C. Raffel,

- mt5: A massively multilingual pre-trained text-to-text transformer, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021, pp. 483–498.
- [10] A. Shvets, L. Wanner, Concept extraction using pointer–generator networks and distant supervision for data augmentation, in: International Conference on Knowledge Engineering and Knowledge Management, Springer, 2020, pp. 120–135.
- [11] E. A. Stathopoulos, A. Shvets, R. Carlini, S. Diplaris, S. Vrochidis, L. Wanner, I. Kompatsiaris, Social media and web sensing on interior and urban design, in: 2022 IEEE Symposium on Computers and Communications (ISCC), IEEE, 2022, pp. 1–6.