# Chapter 18
# Extracting Terminological Concept Systems from Natural Language Text

Dagmar Gromann, Lennart Wachowiak, Christian Lang, and Barbara Heinisch

**Abstract** Terminology denotes a language resource that structures domain-specific knowledge by means of conceptual grouping of terms and their interrelations. Such structured domain knowledge is vital to various specialised communication settings, from corporate language to crisis communication. However, manually curating a terminology is both labour- and time-intensive. Approaches to automatically extract terminology have focused on detecting domain-specific single- and multi-word terms without taking terminological relations into consideration, while knowledge extraction has specialised on named entities and their relations. We present the Text2TCS method to extract single- and multi-word terms, group them by synonymy, and interrelate these groupings by means of a pre-specified relation typology to generate a Terminological Concept System (TCS) from domain-specific text in multiple languages. To this end, the method relies on pre-trained neural language models.

## 1 Overview and Objectives

Domain knowledge is paramount to any specialised communication setting. A structured representation of domain-specific terminology fosters the acquisition of new domain knowledge, the expansion of existing knowledge, and optimises specialised discourse by supporting terminological consistency (Budin 1996). Extracting Terminological Concept Systems from Natural Language Text (Text2TCS) is a pilot project supported by the European Language Grid (ELG) to develop a language technology that automatically extracts a Terminological Concept System (TCS) from domain-specific texts in multiple languages. A TCS is a terminological resource that conceptually structures domain-specific terms and provides hierarchical and non-hierarchical relations between them. Within the context of terminology science, a *term* signifies a domain-specific designation that linguistically represents a domain-specific concept (ISO1087 2019). A *concept* groups terms by meaning, which is

Dagmar Gromann · Lennart Wachowiak · Christian Lang · Barbara Heinisch
University of Vienna, Austria, dagmar.gromann@univie.ac.at, lennart.wachowiak@univie.ac.at, christian.lang@univie.ac.at, barbara.heinisch@univie.ac.at

G. Rehm (ed.), *European Language Grid*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-17258-8_18

generally represented as unique characteristics shared by a set of real-world entities. Once terms have been grouped into concepts based on their synonymous meaning within languages and equivalent meaning across languages, terminology science foresees interrelations of concepts by terminological relations. Such relations are categorised into hierarchical, i. e., generic and partitive, and non-hierarchical, e. g., causal and spatial, relations. For instance, the sentence *COVID causes coughing* can be depicted as a causal relation from the concept that represents the cause *COVID* to the effect concept designated by *coughing*. However, in practice, publicly available terminologies rarely contain any relations, since manually creating them is time- and labour-intensive. While Automated Term Extraction (ATE) methods have proliferated (e. g., Astrakhantsev 2018; Lang et al. 2021), additionally structuring extracted terms by concepts and relations has been neglected. To address this issue, Text2TCS provides a method and tool to extract terms and interrelations between domain-specific synonym sets across languages and domains. The Text2TCS implementation has been integrated and is available on the ELG plattform.[1]

## 2 Methodology

The Text2TCS methodology depicted in Figure 1 builds on a pipeline approach with the following steps: preprocessing, term extraction, relation extraction and post-processing. The pipeline takes domain-specific natural language sentences or text as input and outputs a TCS in the TermBase eXchange (TBX) format and as a concept map. We experimented with several joint term and relation extraction methods, especially relying on pre-trained Neural Machine Translation and Sequence to Sequence models such as mT5 (Xue et al. 2021). However, a pipeline approach relying on fine-tuning XLM-R (Conneau et al. 2020) was finally preferable due to a smaller model size as well as a substantially higher inference speed and performance reliability. In order to fine-tune pre-trained models, training data needs to be available. To this end, two terminologists annotated 51 texts spanning distinct domains from computer science to ecology in English and German with a total of 6,327 terms and 9,460 relations.

### 2.1 Preprocessing

In a first step, the input text's language is detected and it is split into individual sentences. The former relies on the Python library PYCLD2[2] that supports 83 languages. Language detection is required in order to issue a warning in case the input language is unsupported and to indicate the language in the final TBX output file. Furthermore,

[1] https://live.european-language-grid.eu/catalogue/tool-service/8122

[2] https://github.com/aboSamoor/pycld2

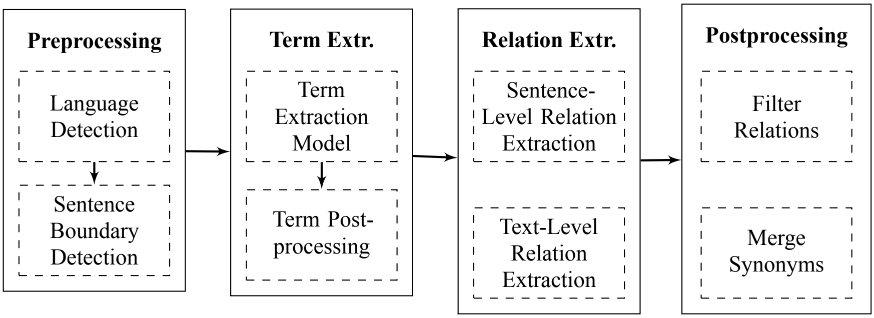

**Fig. 1** Text2TCS extraction pipeline

the detected language is passed on to the sentence boundary detection module that relies on language-specific rules.

Sentence boundary detection is achieved using the rule-based Python module pySBD (Sadvilkar and Neumann 2020), which officially supports 22 languages. This step is required due to limited input length of current neural language models and to allow for a sentence-based relation extraction step. Thus, the pipeline can be sure to support 22 languages (two-digit ISO language codes): am, ar bg, da, de, en, es, el, fa, fr, hi, hy, it, ja, kk, mr, my, nl, ru, pl, ur, zh. However, the term and relation extraction models potentially support up to 100 languages.

## 2.2 Term Extraction

From several distinct experiments with term extraction, which we detail in Lang et al. (2021), the best performing classifies each token of an input sentence separately, utilising the same fully connected layer for all tokens after they have been processed by XLM-R. In term extraction, an established method is (e. g., Hazem et al. 2020) to first generate all possible term candidates from a sequence/sentence and input the candidate together with its context for the model to predict whether it is a term or not. This requires first generating all possible n-grams of a pre-specified length from a text. Instead, the token classification we propose assigns one of three labels to each token in a sequence: `B-T` for beginning of term, `T` for continuation of term, and `n` for not a term (component). For instance, the input sequence “motor vehicle means any power-driven vehicle.” would be labeled as `B-T, T, n, n, B-T, T, n`, extracting the terms “motor vehicle” and “power-driven vehicle”. This approach leads to a substantial reduction in training and inference time compared to previous methods. In XLM-R’s own tokeniser, which we utilise, we noticed an issue with trailing punctuation, e. g. a comma after a term. Thus, we apply an additional cleaning step in which we remove trailing punctuation from a standard punctuation list, unless the punctuation appears multiple times in the term, e. g. “U.S.A.”.

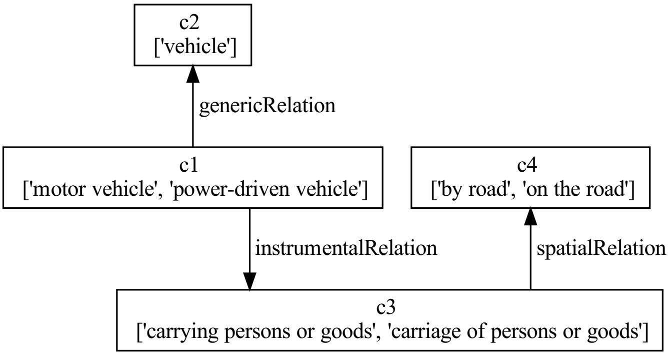
c2
['vehicle']
genericRelation
c1
['motor vehicle', 'power-driven vehicle']
c4
['by road', 'on the road']
instrumentalRelation
spatialRelation
c3
['carrying persons or goods', 'carriage of persons or goods']

**Fig. 2** Example TCS from sequence "motor vehicle means any power-driven vehicle, which is normally used for carrying persons or goods by road or for drawing, on the road, vehicles used for the carriage of persons or goods"

## 2.3 Relation Extraction

Related domain-specific mentions in text can either occur within the same sentence or across sentence boundaries. Thus, two separate models in the pipeline predict relations: a sentence-level and a text-level model. For sentence-level relation extraction, we input a mention pair followed by a contextualising sentence containing both mentions to a fine-tuned pre-trained XLM-R model that predicts a relation taking the relation direction into account (see Wachowiak et al. 2021, for details). We apply our own relation typology of hierarchical relations, i. e., generic and partitive, and non-hierarchical relations, i. e., activity, causal, instrumental, origination, spatial, property, and associative. Generic relations and synonyms frequently occur across sentence boundaries, which is why we additionally train a text-level relation extraction model to detect these two, building on our previous model (Wachowiak et al. 2020) fine-tuning XLM-R. This model takes a mention pair as input and classifies it as a generic relation, synonymy or random, which means no or any other relation. Since predicting relations for individual term pairs drastically impacts inference time, we optimize the pipeline to process multiple term pairs and their context sentence simultaneously.

## 2.4 Postprocessing

In the last step, synonyms predicted on sentence- and text-level are merged into concepts. Furthermore, the relations predicted by the two models are filtered to only include those with high confidence scores and to remove duplicates to provide the final TCS exemplified in Figure 2.

## 3 Evaluation

We evaluated individual steps in the pipeline as well as the overall system on manually TCS-annotated texts in English, German, Spanish, Portuguese, French, Italian, Romanian and Russian as well as on standard datasets, where available, for a better comparison. The term extraction model outperforms previous neural approaches (Hazem et al. 2020) from the TermEval challenge by up to 11.6 F1 score and obtained 74% (Precision: 70%, Recall: 78%) on our dataset. The sentence-level relation extraction model obtained a weighted F1 score of up to 53% (Precision: 56%, Recall: 53%) and the text-level relation extraction model of up to 78% (Precision: 78%, Recall: 77%) on our manually annotated datasets. The sentence-level extraction is also compared to a mixed dataset of the SemEval 2007 Task 4 and SemEval 2010 Task 8 relations, on which the model obtains a weighted F1 score of 87% (see Wachowiak et al. 2021, for details).

## 4 Conclusions and Results of the Pilot Project

Automatically extracting and structuring domain-specific knowledge from text is a challenging task. Text2TCS innovatively fine-tunes pre-trained neural language models in a pipeline approach to first extract terms, second relations on sentence- and text-level, and finally group synonyms. To this end, this pilot project proposed a novel typology of terminological relations. A consistent use of relation types across languages aims to ease the alignment of resulting monolingual TCS across languages. Integrating such an alignment method is future work. At the moment, the method takes terms and relations into consideration, however, text frequently contains (parts of) natural language definitions and their extraction would represent a valuable future addition to the method.

**Acknowledgements** The work described in this article[3] has received funding from the EU project European Language Grid as one of its pilot projects. The computational results presented have been achieved in part using the Vienna Scientific Cluster (VSC).

## References

Astrakhantsev, Nikita (2018). “ATR4S: toolkit with state-of-the-art automatic terms recognition methods in Scala”. In: *Language Resources and Evaluation* 52.3, pp. 853–872.

Budin, Gerhard (1996). *Wissensorganisation und Terminologie: Die Komplexität und Dynamik wissenschaftlicher Informations- und Kommunikationsprozesse*. Vol. 28. Forum für Fachsprachen-Forschung. Gunter Narr Verlag.

[3] https://text2tcs.univie.ac.at

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov (2020). "Unsupervised Cross-lingual Representation Learning at Scale". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault. ACL, pp. 8440–8451. DOI: 10.18653/v1/2020.acl-main.747. URL: https://www.aclweb.org/anthology/2020.acl-main.747.

Hazem, Amir, Mérieme Bouhandi, Florian Boudin, and Beatrice Daille (2020). "TermEval 2020: TALN-LS2N System for Automatic Term Extraction". In: *Proceedings of the 6th International Workshop on Computational Terminology*. Ed. by Béatrice Daille, Kyo Kageura, and Ayla Rigouts Terryn. Marseille, France: ELRA, pp. 95–100.

ISO1087 (2019). *ISO 1087:2019: Terminology work and terminology science – Vocabulary*. Standard. Geneva, CH: International Organization for Standardization.

Lang, Christian, Lennart Wachowiak, Barbara Heinisch, and Dagmar Gromann (2021). "Transforming Term Extraction: Transformer-Based Approaches to Multilingual Term Extraction Across Domains". In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. ACL, pp. 3607–3620. DOI: 10.18653/v1/2021.findings-acl.316.

Sadvilkar, Nipun and Mark Neumann (2020). "PySBD: Pragmatic Sentence Boundary Disambiguation". In: *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*. ACL, pp. 110–114.

Wachowiak, Lennart, Christian Lang, Barbara Heinisch, and Dagmar Gromann (2020). "CogALex-VI Shared Task: Transrelation - A Robust Multilingual Language Model for Multilingual Relation Identification". In: *Proceedings of the Workshop on the Cognitive Aspects of the Lexicon*. Ed. by Rong Xiang, Emmanuele Chersoni, Luca Iacoponi, and Enrico Santus. ACL, pp. 59–64.

Wachowiak, Lennart, Christian Lang, Barbara Heinisch, and Dagmar Gromann (2021). "Towards Learning Terminological Concept Systems from Multilingual Natural Language Text". In: *3rd Conference on Language, Data and Knowledge (LDK 2021)*. Ed. by Dagmar Gromann, Gilles Sérasset, Thierry Declerck, John P. McCrae, Jorge Gracia, Julia Bosque-Gil, Fernando Bobillo, and Barbara Heinisch. Vol. 93. Open Access Series in Informatics (OASIcs). Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 22:1–22:18. DOI: 10.4230/OASIcs.LDK.2021.22.

Xue, Linting, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel (2021). "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*. ACL, pp. 483–498. DOI: 10.18653/v1/2021.naacl-main.41.