Towards Learning Terminological Concept Systems from Multilingual Natural Language Text

Lennart Wachowiak[1]
Centre for Translation Studies, University of Vienna, Austria

Christian Lang
Centre for Translation Studies, University of Vienna, Austria

Barbara Heinisch
Centre for Translation Studies, University of Vienna, Austria

Dagmar Gromann
Centre for Translation Studies, University of Vienna, Austria

Abstract

Terminological Concept Systems (TCS) provide a means of organizing, structuring and representing domain-specific multilingual information and are important to ensure terminological consistency in many tasks, such as translation and cross-border communication. While several approaches to (semi-)automatic term extraction exist, learning their interrelations is vastly underexplored. We propose an automated method to extract terms and relations across natural languages and specialized domains. To this end, we adapt pretrained multilingual neural language models, which we evaluate on term extraction standard datasets with best performing results and a combination of relation extraction standard datasets with competitive results. Code and dataset are publicly available.[2]

2012 ACM Subject Classification Computing methodologies → Information extraction; Computing methodologies → Neural networks; Computing methodologies → Language resources

Keywords and phrases Terminologies, Neural Language Models, Multilingual Information Extraction

Digital Object Identifier 10.4230/OASIcs.LDK.2021.22

Supplementary Material *Software (Source Code and Dataset)*: `https://github.com/Text2TCS/Towards-Learning-Terminological-Concept-Systems`
archived at `swh:1:dir:fff3183f35a3dd332e1d4a2cdc54d21259b4fae2`

Funding This research has been conducted within the Text2TCS project (`https://text2tcs.univie.ac.at/`) funded by European Language Grid (ELG) H2020 No. 825627.

1 Introduction

Terminological inconsistency represents one major source of misunderstanding in specialized communication. One vital measure to counteract such inconsistency is the creation of a TCS that represents concepts, their terms and interrelations. Thereby, it can be ensured that different parties in a communication, such as medical, political, and news teams in times of crisis, consistently refer to phenomena by utilizing the same words. Several approaches to automatically extract domain-specific terms, i.e., single- and multi-word sequences, from natural language text exist. Such methods rely on frequency-based to Wikipedia-link-based Automated Term Extraction (ATE) approaches [4]. ATE is further distinguished depending on whether it is performed on document or corpus level. However, to the best of our knowledge, no approaches to extract a full terminological concept system from multilingual texts have been proposed.

[1] Corresponding author
[2] `https://github.com/Text2TCS/Towards-Learning-Terminological-Concept-Systems`

3rd Conference on Language, Data and Knowledge (LDK 2021).
Editors: Dagmar Gromann, Gilles Sérasset, Thierry Declerck, John P. McCrae, Jorge Gracia, Julia Bosque-Gil, Fernando Bobillo, and Barbara Heinisch; Article No. 22; pp. 22:1–22:18
OpenAccess Series in Informatics
OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

A TCS groups synonyms and equivalents across languages into a single concept and interrelates these concepts with a set of prespecified relations. A major distinction is made between hierarchical, i.e., generic and partitive, and non-hierarchical, e.g. activity and ownership, relations. While to the best of our knowledge there are no approaches for learning a TCS from text, neighboring fields can provide inspiration for the task at hand. For instance, entity linking represents the task of identifying and interlinking named entities based on information provided in text (e.g. [33]) and ontology learning also requires term and relation extraction (e.g. [31] build on deep learning).

In this work, we rely on recent developments of deep learning and especially the recent success of large pretrained multilingual language models. In our approach we split the task of learning a TCS from text into two sentence-level steps: 1) term extraction, and 2) relation extraction. We rely on adaptations of the pretrained multilingual language model XLM-RoBERTa (XLM-R) [9] for both steps and connect them in a pipeline. The first step reads the document sentence by sentence and assigns each word with one of three tags: term, term continuation, not a term. In the second step and with a different adaptation of XLM-R, we identify relations between terms building on a predefined set of hierarchical and non-hierarchical terminological relations.

Given the resulting information we automatically learn from text, this approach contributes to the topic of knowledge graphs as well as deep learning for Linguistic Linked Open Data (LLOD). Since XLM-R is highly multilingual, trained on 100 different languages, TCSs can be learned with the proposed approach from texts in any of those languages. Nevertheless, very few datasets for evaluating multilingual TCSs exist. For term extraction, we train and evaluate the system on the TermEval 2020 dataset [34] in English, French, and Dutch across four specialized domains as well as the English ACL RD-TEC 2.0 dataset [32]. For relation extraction, we rely on a combination of SemEval datasets [15, 20], a WCL hypernymy dataset [28], and manually annotated data we created, all of which are only available in English. To represent the resulting TCS, we currently rely on the ISO standard TermBase Exchange (TBX) format, however, the resulting information could be serialized in any adequate format. Methods for hosting terminological resources as LLOD have been proposed before (e.g. [11]).

In the next section, we present a brief theoretical introduction to a TCS and terminology, followed by an introduction to language models in Section 3. Section 4 details the data utilized to train and evaluate the proposed approach. Section 5 details the TCS learning method as well as the individual steps thereof, the results of which are presented in Section 6. We discuss the results in Section 7 and present related work in Section 8 prior to some concluding remarks.

2 Terminological Concept Systems

This section provides a brief overview of the field of terminology and TCS. It also states the relation typology utilized for our TCS learning approach.

2.1 Term, Concept and Terminology

Terminology can only be understood within the framework of specialized language or language for special purposes, which is defined as "**natural language** [...] used in communication between experts in a **domain** [...] and characterized by the use of specific linguistic means of expression" (emphasis in original) [1]. Thus, terminology refers to a set of concepts and their designations in a specialized field of knowledge (a language for special purposes). In

terminology studies, different schools of thought exist and we will follow the so-called General Theory of Terminology, where concepts and terms are differentiated, wherein concepts are considered abstractions of a set of physical or abstract entities and terms are their designation by linguistic means [3].

A designation can refer to a single- or multi-word term, a named entity, a symbol or even a formula. When talking about term extraction, in general extracting named entities and symbols is implicitly included. Concepts are rather vaguely referred to as units of thoughts and knowledge, however, we treat them as structuring means for synonymous terms. A concept system refers to the organization of concepts, and thereby also of knowledge.

Many ISO standards are based on this school of thought in terminology. The ISO standards address topics that range from the definition of terminology and terminology management to the representation of terminology in terminological databases. Among these standards is the TermBase eXchange (TBX) standard [2] that defines an industry representation format for exchanging terminological resources detailed below.

2.2 Relation Types

After identifying concepts, they can be analyzed and modeled by means of concept systems. Concept relations describe the link between different concepts. In the literature on terminology, different models for describing concept relations have been proposed, at times with a very large typology of relation types (e.g. [30]). On the highest level, relation types are generally classified into hierarchical and non-hierarchical relations.

Hierarchical relations connect a superordinate with a subordinate concept and are either generic or partitive. Generic relations exist "between two concepts when the intension of the subordinate concept includes the intension of the superordinate concept plus at least one additional delimiting characteristic" [3], such as "furniture" which serves as the superordinate concept for "desk". A lexical manifestation of this relation could, for instance, be "desk is a kind of furniture". Partitive relations exist when the superordinate concept relates to the entire object and the subordinate concept refers to its parts. For example, "root", "branch" and "stem" are parts of the superordinate concept "tree", or linguistically described a "stem is part of a tree".

Non-hierarchical relations are called associative in ISO 1087 [1], however, in the typology we adopt, an associative relation represents a thematic connection between two concepts that is not further specified. The number of non-hierarchical relations in the ISO 1087 standard [1] is rather small – sequential as superordinate to spatial, temporal and causal relations – and has been criticized for being inconsistent and ambiguous. Fortunately, a variety of relations have been proposed by different authors (e.g. [30]). In ontology learning and knowledge graph generation non-hierarchical relations also play a crucial role. However, the relation types generally vary significantly from one ontology or knowledge graph to another. While in the future we seek to map our typology to existing standard LLOD resources, for now we adopt relation types established in the terminology community and a consistent typology across domains and languages.

The relation types used in this study are derived from a literature review and were adapted to the needs of this research. The objective is to map semantic relations to this prespecified typology in order to ease the alignment between different TCSs resulting from our method across domains and languages, which consists of: *generic relation* (is a kind of, e.g. table is a kind of furniture) and *partitive relation* (is part of, e.g. roots are part of a tree), several non-hierarchical relations were included, that is, *spatial relation* (for objects

and their location, e.g. avalanche and mountain), *temporal relation* (for objects and their time or sequences, e.g. production and consumption), *causal relation* (for causes and their effect, e.g. accident and injury),

- Hierarchical relations:
  - *generic relation* the intension of the subordinate concept includes the intension of the superordinate concept plus one additional characteristic, e.g. "table is a kind of furniture"
  - *partitive relation* the subordinate concepts are parts of the superordinate concept, e.g. "roots are part of a tree"
- Non-hierarchical relations:
  - *activity relation* connects actors with an activity or an activity with its entity, e.g. "teacher activity schoolchildren" where the activity can be teaching,
  - *causal relation* connects causes and their effect, e.g. "accident causes injury",
  - *instrumental relation*: connects instruments and their use, e.g. "coffee machine instrumental coffee" since the former is utilized to make the latter,
  - *origination relation* connects an entity with its origin, e.g. "car origination factory" since the car originates from a factory,
  - *spatial relation* connects an entity with its location, e.g. "avalanche spatial mountain" since the former is located on the latter
  - *associative relation* provides a generic thematic connection between concepts, e.g. "lecturer associative education" since both are thematically associated to each other.

An associative relation can serve to model a connection between two loosely related concepts to which none of the other relation types applies. Apart from the symmetric associative relation, all relations are directed, e.g. for the instrumental relation, the direction is from instruments to their use and the partitive relation is directed from parts to whole. We initially treat synonymy as a symmetric relation in the relation extraction step, even though in the final output synonymy is not represented as relation, but as a set of synonyms to form a concept called terminological entry.

2.3 Representation in TBX

ISO 30042 [2] defines a framework to represent terminological data in a structured format called TBX, which aims at facilitating the exchange of terminological data for different purposes, including analysis, representation and dissemination. The users of TBX files include terminologists, translators or technical writers on the one hand, and computer applications, such as computer-assisted translation tools and authoring software, on the other. It is a flexible format that allows for user-defined relation types. As the de-facto standard for terminological resource representation in industry and academia, we opted for TBX as our initial output format, but intend to accommodate RDF directly and not only by way of conversion of TBX to RDF [11] in the near future.

3 Language Models

Most of the recent progress in natural language processing can be traced back to transformer-based pretrained language models. This type of transfer learning utilizes deep neural networks based on the transformer architecture [38]. In a first stage called pretraining, the network learns to predict a masked word given its context, a task for which large amounts of training data are available. In the second stage called fine-tuning, the pretrained network is used

again and is trained for a specific task like classification by adding additional layers on top of the model, while making use of the previously learned rich language representations. A frequently used English language model is BERT [10], for which also multilingual variants exist that have been pretrained on corpora in multiple languages, e.g. multilingual BERT or XLM-R [9]. XLM-R uses the enhanced training paradigm introduced by RoBERTa [26] while being pretrained on a CommonCrawl dataset in 100 different languages. Due to the pretraining, multilingual models can be fine-tuned in one language and show strong zero-shot performances in another language, for which no training samples were provided during fine-tuning.

4 Data

The data used for training and evaluation was compiled from multiple datasets. In order to effectively use the data to train our model for the specified tasks, some pre-processing as well as manual creation of additional data was necessary.

4.1 Term Extraction Data

For term extraction, we used the Annotated Corpora for Term Extraction Research (ACTER)[3] that was also used in the TermEval 2020 challenge [34]. This provided us with hand-annotated data and a good baseline to evaluate our systems. The data comes in the form of raw text documents divided into four domains (corruption, dressage/equitation, wind energy, heart failure) and a single document per domain containing all terms that have been identified in the text documents by language experts. The terms are provided with the same surface-form as they appear in the texts, so each term-list may contain several morphological variations of a term. All terms are provided in lower-case. Additionally, the data is provided in three languages, namely English, French, and Dutch. At the time of writing, the most recent version of the dataset was version 1.4, which did not provide inline annotations. Since our training is performed on sentences, we needed to annotate each sentence in the provided text documents with the terms from the corresponding term document.

To annotate the texts, we first split the documents into sentences using the spaCy sentencizer [21] and tokenized each sentence using sacremoses[4], a tokenizer written in Python. Subsequently, each individual sentence was annotated with terms from the term document of the corresponding domain. Only terms that had a full match with any word or word sequence composing each sentence were annotated. This way it was possible to create an inline annotation from the raw data. In order to allow a comparison with the TermEval results we followed the train-val-test-split of the TermEval 2020 challenge and used the corruption and wind energy domains as training, the dressage (equitation) domain as validation, and the heart failure domain as test data set. With this train-val-test-split, around 10,000 sentences per language were used for supervised training, while the test set contained approximately 2,000 sentences. The exact word and term count per language is represented in Table 2.

We also trained separate models using the ACL RD-TEC 2.0 dataset [32] to verify if our approach would work on other datasets. The ACL RD-TEC 2.0 dataset provides high-quality inline annotations of 471 scientific abstracts by two human annotators. In total more than 2,200 sentences were annotated with 6,818 terms. Annotator 1 annotated 900 sentences and

[3] https://github.com/AylaRT/ACTER
[4] https://github.com/alvations/sacremoses

Annotator 2 1,301 sentences. Since no split is proposed by the dataset authors, we split the data into 60% training data, 20% validation data, and 20% test data for each of the two annotators respectively. This resulted in 540 training sentences and 180 val/test sentences for Annotator 1 and 780 training sentences and 260/261 val/test sentences for Annotator 2 respectively. Since the dataset is inline annotated, the terms from each sentence could be easily extracted with an XML Parser. Additionally, unlike TermEval, capitalization of terms is maintained both for the training and validation/test data.

4.2 Relation Extraction Data

For relation extraction, we combined training data from two SemEval tasks to obtain more training examples with a higher diversity of relation types: SemEval 2007 Task 4 [15] and SemEval 2010 Task 8 [20]. We then mapped the relation types of these datasets to the relation types defined for our TCS (see Section 2.2). Since these two datasets lack generic relations, we additionally utilized the manually created WCL 1.2 dataset [28] and automatically annotated the terms in the ACTER dataset texts with generic relations and synonyms, which were also not represented in the other datasets, by relying on WordNet relations. Furthermore, we extracted acronyms and their long unabbreviated forms as synonyms from the ACTER texts by adapting the regex-based acronym extraction method proposed by Azimi et al. [6]. We further added to the data by manually annotating around 200 acronym-term pairs from ACTER with relations other than synonymy and another 271 sample sentences from the Common Crawl News Corpus[5] with term pairs that show no relation at all, i.e., negative samples to be classified as "none". All samples of the resulting relation dataset are in English. Statistics regarding the relation type distribution can be seen in Figure 1.

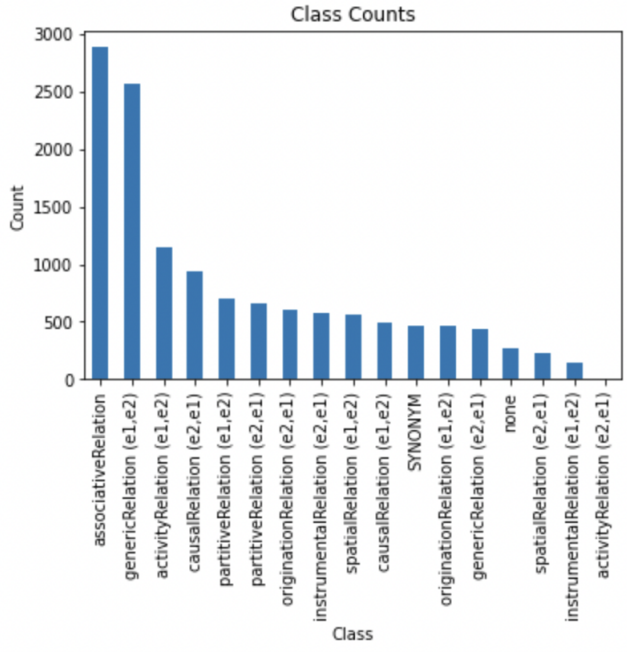

Figure 1 Number of samples for each relation type.

[5] https://commoncrawl.org/2016/10/news-dataset-available/

For the data originating from SemEval we use the original train-test split, while from the additional data we take 20% for testing. An additional 20% from all resulting training data is used for validation.

5 Method

We will first describe the overall architecture of the proposed TCS learning pipeline. Subsequently, we introduce our term extraction and relation extraction models.

5.1 Architecture

The TCS extraction pipeline, as shown in Figure 2, takes text documents as input and outputs terminologies in the TBX format as well as in the form of a graph visualization, an example of which is shown in Figure 3. Due to input length restrictions of current language models, terms as well as their relations are extracted on sentence-level basis. Thus, as the first and only preprocessing step, the input document is split into sentences using the rule-based sentence segmentation component provided by the software library spaCy [21].

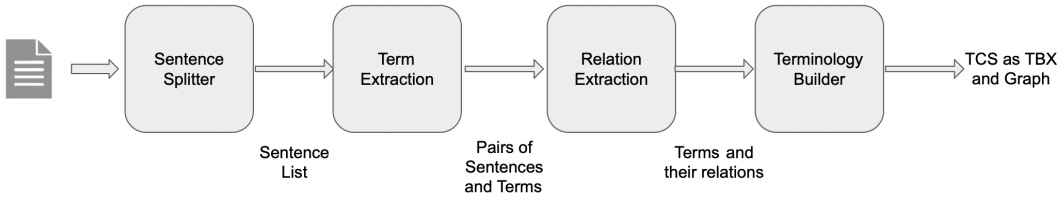

Figure 2 TCS learning pipeline.

In a second step, terms are extracted from each sentence using the neural network described in Section 5.2. After the term extraction step, we end up with pairs of sentences and the terms they contain according to the model.

Based on this data, all possible pairs of terms are computed per sentence. These pairs together with their respective context, i.e., the sentence which contains them, are then fed one after the other to the second neural network, described in Section 5.3, which outputs whether or not there is a relation between the terms and, if so, which relation type exactly.

As the fourth step, these term pairs and their corresponding relations are used to create a terminological concept system. Therefore, terms with synonym relations are merged into concepts with a unique identifier. The extracted relations, which at this point are still between specific terms, are mapped to the newly created concepts. Through this process it is possible that self-referential relations as well as duplicate relations are created, which are subsequently deleted. Moreover, it is possible that there are multiple different relations between the same two concepts, however, only one is represented in the final TCS. To provide more insight into this process to the final user, we show the extraction network's classification confidence in the output. Lastly, the resulting TCS is represented in the TBX format and as a graph utilizing the Graphviz library [13].

5.2 Term Extraction Model

For term extraction, we take advantage of the multilingual pretrained language model XLM-R in its base size made available by the transformers library [40]. The input provided to the model consists of full sentences and is based on the data described in Section 4.1. The output

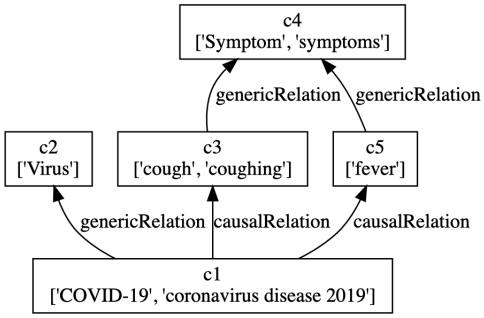

Figure 3 Example of a possible visualization with Graphviz.

labels are given on a word basis, i.e., each word is either tagged as "B-T" (beginning of a term), "T" (continuation of a term), or "n" (no term). For instance, the labels for the sentence "We meta-analyzed mortality using random-effect models" are "n", "B-T", "B-T", "n", "B-T", "T". For classification into these three classes we use a single fully connected layer which uses the representations created by XLM-R for the individual words as input. Since XLM-R tokenizes the input on a subword level, we obtain labels for these subword units which have to be mapped back to the original input. Training was conducted using the Adam optimizer with a learning rate of 2e-5, a batch size of 8, and a validation every 100 steps allowing us to load the best performing model at the end of the training procedure, which consisted of 10 epochs overall.

5.3 Relation Extraction Model

As with term extraction, we fine-tune the pretrained XLM-R for the relation extraction task. The data used is described in Section 4.2. The input of the model consists of an entity pair followed by a contextualizing sentence containing both entities, for instance, "cough. Covid-19. The cough was caused by Covid-19." The model classifies the representation of the whole sequence created by XLM-R as input with a fully connected layer into one of the relations presented in Section 2.2. For directional relations two classes are available, so that the given example input would, for instance, be classified as `causalRelation(e2,e1)`. The model was trained for 9 epochs utilizing the Adam optimizer with a learning rate of 2e-5 and a batch size of 32.

6 Results

Since no datasets for a full TCS evaluation are available as of yet, we evaluate our model on the described datasets separately and present the results below.

6.1 Term Extraction

For the term extraction step, we evaluated our model in comparison to the best preforming models of the TermEval 2020 shared task in terms of precision, recall and F1 score as shown in Table 1. These metrics are calculated on the basis of the available annotation in the original ACTER dataset, where we opted for the more comprehensive list of terms including named entities. Since different combinations of training and test languages might have an impact on the overall performance, we report on these combinations in Table 1. As in TermEval 2020 we use the heart failure domain as hold-out test set. The baseline for English and

French is provided by [19], who used monolingual neural language models to predict whether a given phrase is a term provided some context. Thus, other combinations as tested with our multilingual model are not available with monolingual models. The baseline for Dutch is provided by a bidirectional LSTM with GLOVE[6]. The overall best results are marked in bold for each test language. For the ACL RD-TEC 2.0 corpus no baseline is available and the data split into 60% training, 20% validation, and 20% test data was chosen by us as no split was suggested by the authors of the dataset. The results are also available in Table 1.

Table 1 Test set results of our term extraction model on two different datasets evaluated on different langauge combinations.

Dataset	**Training**	**Test**	**Token Classifier**			**Previous SOTA**		
			Prec	Rec	F1	Prec	Rec	F1
TermEval 2020	EN	EN	54.9	62.2	**58.3**	34.8	70.9	46.7
TermEval 2020	FR	EN	56.7	36.2	44.2			
TermEval 2020	NL	EN	55.3	61.8	**58.3**			
TermEval 2020	ALL	EN	54.4	58.2	56.2			
TermEval 2020	EN	FR	65.4	51.4	**57.6**			
TermEval 2020	FR	FR	68.7	43.0	52.9	44.2	51.5	48.1
TermEval 2020	NL	FR	62.3	48.5	54.5			
TermEval 2020	ALL	FR	49.4	55.3	55.0			
TermEval 2020	EN	NL	67.9	71.7	**69.8**			
TermEval 2020	FR	NL	69.2	55.2	61.4			
TermEval 2020	NL	NL	71.4	67.8	69.6	18.9	18.6	18.7
TermEval 2020	ALL	NL	70.0	65.8	67.8			
ACL (An.1)	EN	EN	74.4	77.2	75.8			
ACL (An.2)	EN	EN	80.1	79.3	80.0			

It can be seen from Table 1 that our solution outperforms the TermEval 2020 baseline in all three languages. However, it is interesting to see that mixed training and test languages achieve best results. As a matter of fact, the model trained on English achieved best results not only when tested on English, but also when tested on French and Dutch. When the test language was English, training on Dutch achieved equivalent results to training on English. A general assumption would be that training on the language that is being tested or training on all available languages should perform best, an assumption that could not be confirmed in this experiment. Furthermore, a substantial difference in recall behavior can be observed from one and the same model, which can also be observed with the monolingually pretrained baseline models even though they achieve a competitive or even higher recall. This suggests differences in term type across the three languages. This observation can also be confirmed in the validation set performance reported in Table 2, where French as training and validation set performs significantly worse. Validation results also show some performance differences to the test performances.

The models trained on the ACL RD-TEC 2.0 corpus show an even stronger performance with an F1 score of 75.8 and 80.0 for Annotator 1 and 2 respectively. Moreover, the scores for precision and recall of the two resulting models are nearly perfectly balanced. The validation scores, reported in Table 2, are consistent with the test scores.

[6] No system description paper was submitted for this approach after participation in the challenge.

Table 2 Train, validation and test split by word count (W_{lang}) and term count (T_{lang}) per language (left) and validation performance of token classifier (right).

	Train	Val	Test
W_{en}	97,145	51,470	45,788
T_{en}	2,708	1,575	2,585
W_{fr}	106,792	53,316	46,751
T_{fr}	2,185	1,183	2,423
W_{nl}	96,887	50,882	47,888
T_{nl}	2,540	1,546	2,257

Training	EN Val	FR Val	NL Val
EN (TermEval)	55.6	45.3	60.5
FR (TermEval)	41.9	33.6	49.6
NL (TermEval)	54.6	47.7	57.8
ALL (TermEval)	50.0	40.4	51.5
EN (ACL An.1)	75.5	/	/
EN (ACL An.2)	79.3	/	/

In terms of term type, the models trained on TermEval 2020 are able to handle acronyms well, which might be due to the fact that much of the training data was based on rather technical documents like scientific abstracts. However, if acronyms are part of the term, e.g. “LV strain rate”, there was a high number of false negatives. Equally apostrophes in named entities represented a challenge, e.g. “Chaga’s disease”. We could observe a tendency of the model to split particularly long multi-word sequences (more than five words), e.g. “resynchronization reverses remodeling in systolic left ventricular dysfunction”. We made similar observations when manually evaluating the model trained and tested on Annotator 1 of ACL RD-TEC 2.0 dataset. While performance on acronym extraction was generally good, if acronyms were part of a term, it likely resulted in a false negative. It can be observed, that false positives often correlate with the false negatives, as the model extracts only parts of the original term or splits the longer terms, e.g. “LRE project SmTA double check” is extracted as “LRE” and “SmTA double check”. It was especially noticeable that the model had difficulties extracting terms containing their acronym or the expansion of an acronym in parentheses, e.g. “machine translation (MT) systems”. This issue also extended to other terms containing parentheses, such as “document descriptors (keywords)”. In fact, not a single example of terms containing parentheses was extracted. Similar to the model trained on TermEval 2020 data, the model trained on the ACL RD-TEC 2.0 data showed a general tendency for extracting shorter terms, with the largest group of false negatives being terms composed of four or more words (41 out of 124 examples).

The model trained on the TermEval 2020 dataset turned out to be highly efficient in terms of training time. Looking at the epochs required to reach the best score on the validation set, we can observe that in most cases the token classifier model requires not even a single training epoch. Training with the English dataset required 300 steps with a full epoch consisting of 432 steps. The model trained on French was the only model with its best performance being reached during the second epoch after 700 steps, while a full epoch consists of 437 steps. The model trained on Dutch performed best after 400 steps while one epoch takes 553 steps. The multilingual model converged the quickest needing only 200 steps whereas a full epoch consists of 1,421 steps. The models trained on the ACL RD-TEC 2.0 dataset need more epochs and achieve their highest scores after 3 and 5 epochs respectively. However, due to the lower training set size of the ACL RD-TEC 2.0 corpus this also corresponds to less than 500 steps, thus, being similar with the training times reported for the model trained with TermEval 2020 data.

6.2 Relation Extraction

The trained model achieves a weighted averaged F1 score of 87% with a precision of 87% and a recall of 87% on the hold-out test set. The confusion matrix in Figure 4 and Table 3 show which classes were learned best. Only the activity relation from entity 2 to entity 1 was not

Table 3 Test performance of the relation classifier and number of test samples.

Relation Type	Precision	Recall	F1	Test samples
synonymy	0.85	0.76	0.80	89
activityRelation (e1,e2)	0.93	0.97	0.95	293
activityRelation (e2,e1)	0.00	0.00	0.00	2
associativeRelation	0.90	0.92	0.91	783
causalRelation (e1,e2)	0.90	0.95	0.92	135
causalRelation (e2,e1)	0.92	0.91	0.91	222
genericRelation (e1,e2)	0.90	0.93	0.92	533
genericRelation (e2,e1)	0.46	0.41	0.43	91
instrumentalRelation (e1,e2)	0.72	0.68	0.70	34
instrumentalRelation (e2,e1)	0.85	0.88	0.86	144
none	0.69	0.44	0.54	70
originationRelation (e1,e2)	0.83	0.89	0.86	116
originationRelation (e2,e1)	0.84	0.83	0.83	165
partitiveRelation (e1,e2)	0.90	0.85	0.87	176
partitiveRelation (e2,e1)	0.77	0.77	0.77	168
spatialRelation (e1,e2)	0.90	0.91	0.91	169
spatialRelation (e2,e1)	0.90	0.82	0.86	44

learned at all given the current training data as the class consists of overall less than 10 data points. Activity relations are usually directed from actor to activity, which was also the case in our dataset, i.e., the actor was mostly mentioned first (e1) and the activity second (e2), with less than 10 exceptions where the actor was mentioned second. The only other relations with an F1 score lower than 0.7 are the none-relation and the generic relation from e2 to e1, which also can be traced back to relatively small amounts of training data as well as a high confusion with the same relation in the opposite direction in case of the generic relation. For the synonymy relation, the classification of synonym pairs including an acronym works well, while the relation of two longer sequence (not shortened) pairs is often confused as a generic relation. Many of the other relations, especially those supported by large amounts of training data, achieve high F1 scores of up to 95%. Furthermore, we can observe very balanced precision and recall scores for all relations.

7 Discussion

Currently, the proposed pipeline fully operates on a sentence-level. In the future we plan to extend the architecture so that the model can extract relations which span over the whole document. This could be achieved by models trained on an appropriate dataset containing such relations. However, currently such datasets are rare and available ones are either domain-specific, very small, and/or focus on named entities [25, 42]. Another option to extract document-wide relations is to add a model to the pipeline which makes predictions about the relation between two words independent from any context in which they appear, something for which, for instance, approaches for hierarchical relations exist [39]. Such a model could be applied to all possible term-pairs, however, due to the missing contexts only limited effectiveness can be expected.

A problem of pipeline approaches is that errors from earlier pipeline steps are propagated to the later components. In the case of the TCS extraction pipeline, wrongly extracted terms are sent to the relation extraction component which tries to establish a relation to

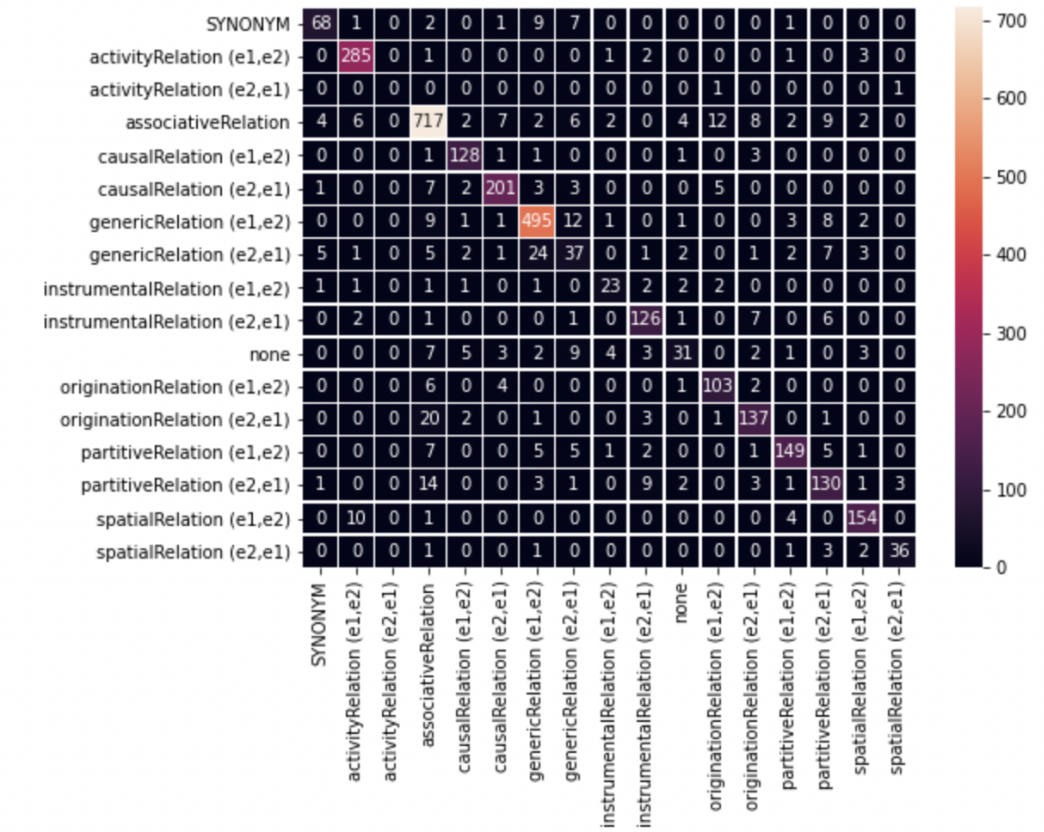

Figure 4 Confusion Matrix for the relation extraction model on the test set.

other terms. This problem can potentially be solved by joint models, which learn to extract terms and their relations together, as was already done in the case of named entities and very limited domain-specific entities and their relations (e.g. [22, 29]). Such models, however, require datasets that annotate terms and their corresponding relations in the same texts, which is something that is currently not available, but something that we are aiming to make available in the future.

8 Related Work

Since the proposed pipeline to automatically learn TCS from text relies on two intermediate steps, term and relation extraction, as well as their combination, we separate the related work into the individual steps as well as approaches joining both steps.

8.1 Term Extraction

An initial classification of ATE methods into statistical, linguistic or hybrid has recently been refined to methods based on term occurrence frequencies, occurrence contexts, domain-specific corpora combined with general language corpora, topic modeling, and those utilizing Wikipedia (see [4] for an overview). Methods are additionally categorized by the type of context, i.e., corpus-level (e.g. [4, 45]) and document-level (e.g. [37]) settings. However, neural ATE methods frequently operating on sentence-level cannot be easily accommodated by these classifications.

The approach most closely related to ours, which also provided our baseline [19], utilized RoBERTa [26] for English and CamemBERT [27] for French and won the TermEval 2020 challenge. In their work, pretrained language models clearly outperformed a classification method based on a variety of features, such as statistical descriptors. They, however, train their model using pairs of context sentences and possible candidate terms, which are based on all possible n-grams contained in a given context sentence, a procedure much slower than our proposed token-level classifier. A recently published approach [36] relies on LSTM, GRU and BERT embeddings and achieves high F1 scores for ATE of Lithuanian terms in the cybersecurity domain. Several approaches build on word embeddings to perform ATE on specific domains, such as medicine (e.g. [7]), or to separate general-language from domain-specific embeddings [18]. In contrast, our model performs ATE on four domains and in three languages utilizing a pretrained language.

8.2 Relation Extraction

Relation extraction describes the supervised task of classifying a relation given two entities and a context. Most work and datasets in the field focus on sentence-level relation extraction [15, 20, 44] with only some exceptions providing relations over longer text spans [42]. Current state-of-the-art approaches for such datasets usually rely on either transformer-based architectures [41, 47] or graph-based neural networks [17, 43].

8.3 Joint Term and Relation Extraction

While to the best of our knowledge no approaches exist to automatically extract terminological concept systems from text, there is an entire research field on connecting terminological information with ontologies, thereby providing relational information to terms. Methods for modeling terminological information as ontologies are generally called terminological ontologies (e.g. [24]). Approaches that model terminological information in relation to ontologies are generally called ontology-terminology models (e.g. [35, 16]). One approach in this direction that is probably most closely related to the one proposed in this paper is TERMINAE [5], a platform that utilizes traditional NLP tools and methods to propose term candidates and relations to users for manual editing by building on terminology engineering principles and findings from ontology learning. While a very interesting method and platform, the approach neither commits to a specific typology of relations nor seeks to provide a fully automated solution. Thus, to broaden the scope of this discussion on related work and consider related fields that directly inspire this joint task, we will discuss two additional research directions. First, we present approaches on entity and relation extraction. Second, we relate to selected ontology learning approaches.

Joint entity and relation extraction (e.g. [46]) is the task of identifying named entities in text and detecting their semantic relations. Approaches to this task range from joining a bidirectional LSTM for term extraction with a CNN for relation extraction [46] to utilizing a Graph Convolutional Network [12]. This idea of joining recurrence and convolution operations is taken up again by Geng et al. [14]. The approach probably most similar to ours is that of Quiao et al. [33] who utilize BERT for joint entity and relation extraction in the agricultural domain. However, our approach has been applied across several domains and languages. In addition, named entities are a subcategory of single- and multi-word terms, where the latter is considerably more challenging. The type of relation is also frequently restricted to lexical-semantic relations, such as synonymy or hypernymy, specific semantic relations, such as the temporal relation, or information in a specific domain, e.g. agriculture.

Ontology learning (e.g. [31, 8]) is the task of automatically extracting knowledge from text, starting with terms which are organized to form concepts, their interrelations which are organized hierarchically and non-hierarchically, and finally axioms. Petrucci et al. [31] utilize Neural Machine Translation (NMT) on a synthetically generated dataset to learn Description Logic formulas from natural language sentences. In contrast, our approach operates on non-synthetic, real-life datasets. Few other approaches utilize deep learning for ontology learning (see [23] for an overview).

9 Conclusion

As a first step to approach fully automated TCS learning from multilingual text, we propose adaptations of pretrained language models to perform term and relation extraction in a pipeline approach. While a multilingual, cross-domain dataset for term extraction exists, we had to accumulate and extend several relation extraction datasets to accommodate a common terminological relation typology. Term extraction results substantially outperform previous results and the relation extraction model achieves competitive results, even though no baseline comparison was available for exactly these relation types.

As a next step we will manually create a full evaluation dataset for TCS across domains and languages to provide a better evaluation scenario for the proposed approach. Additionally, the model currently exclusively extracts information from sentences, whereby several global relations beyond the sentential level will be lost, especially synonymy and generic relations. We thus currently evaluate methods for achieving document-level TCS learning. Lastly, we will extend the set of covered relations by including data for temporal, property, and ownership relations.

References

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

2 ISO 30042:2019. Management of terminology resources – TermBase eXchange (TBX). Standard, International Organization for Standardization, Geneva, CH, 2019.

3 ISO 704:2009. Terminology work – Principles and methods. Standard, International Organization for Standardization, Geneva, CH, 2009.

4 Nikita Astrakhantsev. ATR4S: toolkit with state-of-the-art automatic terms recognition methods in scala. *Language Resources and Evaluation*, 52(3):853–872, 2018.

5 Nathalie Aussenac-Gilles, Sylvie Despres, and Sylvie Szulman. The terminae method and platform for ontology engineering from texts, 2008.

6 Sasan Azimi, Hadi Veisi, and Reyhaneh Amouie. A method for automatic detection of acronyms in texts and building a dataset for acronym disambiguation. In *2019 5th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS)*, pages 1–4. IEEE, 2019. doi:10.1109/ICSPIS48872.2019.9066084.

7 Matthias Bay, Daniel Bruneß, Miriam Herold, Christian Schulze, Michael Guckert, and Mirjam Minor. Term extraction from medical documents using word embeddings. In *4th IEEE Conference on Machine Learning and Natural Language Processing (MNLP 2020)*. IEEE Computer Society, 2020. URL: http://www.wi.cs.uni-frankfurt.de/webdav/publications/TLDIA_Paper_IEEE_CRC.pdf.

8 Philipp Cimiano and Johanna Völker. text2onto. In *International conference on application of natural language to information systems*, pages 227–238. Springer, 2005.

9 Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.747.

10 Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi:10.18653/v1/N19-1423.

11 Maria Pia di Buono, Philipp Cimiano, Mohammad Fazleh Elahi, and Frank Grimm. Terme-à-LLOD: Simplifying the conversion and hosting of terminological resources as linked data. In Maxim Ionov, John P. McCrae, Christian Chiarcos, Thierry Declerck, Julia Bosque-Gil, and Jorge Gracia, editors, *Proceedings of the 7th Workshop on Linked Data in Linguistics (LDL-2020)*, pages 28–35, Marseille, France, May 2020. European Language Resources Association. URL: https://www.aclweb.org/anthology/2020.ldl-1.5.

12 Tsu-Jui Fu, Peng-Hsuan Li, and Wei-Yun Ma. GraphRel: Modeling text as relational graphs for joint entity and relation extraction. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1409–1418, Florence, Italy, July 2019. Association for Computational Linguistics. doi:10.18653/v1/P19-1136.

13 Emden R. Gansner and Stephen C. North. An open graph visualization system and its applications to software engineering. *Software – Practice and Experience*, 30(11):1203–1233, 2000.

14 Zhiqiang Geng, Yanhui Zhang, and Yongming Han. Joint entity and relation extraction model based on rich semantics. *Neurocomputing*, 429:132–140, 2021. doi:10.1016/j.neucom.2020.12.037.

15 Roxana Girju, Preslav Nakov, Vivi Nastase, Stan Szpakowicz, Peter Turney, and Deniz Yuret. SemEval-2007 task 04: Classification of semantic relations between nominals. In Eneko Agirre, Lluís Màrquez, and Richard Wicentowski, editors, *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 13–18, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL: https://www.aclweb.org/anthology/S07-1003.

16 Dagmar Gromann. A model and method to terminologize existing domain ontologies. In *Terminology and Knowledge Engineering 2014*, pages 10–p, 2014.

17 Zhijiang Guo, Yan Zhang, and Wei Lu. Attention guided graph convolutional networks for relation extraction. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 241–251, Florence, Italy, July 2019. Association for Computational Linguistics. doi:10.18653/v1/P19-1024.

18 Anna Hätty, Dominik Schlechtweg, Michael Dorna, and Sabine Schulte im Walde. Predicting degrees of technicality in automatic terminology extraction. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2883–2889, Online, July 2020. Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.258.

19 Amir Hazem, Mérieme Bouhandi, Florian Boudin, and Beatrice Daille. TermEval 2020: TALN-LS2N system for automatic term extraction. In Béatrice Daille, Kyo Kageura, and Ayla Rigouts Terryn, editors, *Proceedings of the 6th International Workshop on Computational Terminology*, pages 95–100, Marseille, France, 2020. European Language Resources Association. URL: https://www.aclweb.org/anthology/2020.computerm-1.13.

20 Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In Katrin Erk and Carlo Strapparava, editors, *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 33–38, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL: https://www.aclweb.org/anthology/S10-1006.

21 Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength Natural Language Processing in Python, 2020. doi:10.5281/zenodo.1212303.

22 Bin Ji, Jie Yu, Shasha Li, Jun Ma, Qingbo Wu, Yusong Tan, and Huijun Liu. Span-based joint entity and relation extraction with attention-based span-specific and contextual semantic representations. In Donia Scott, Nuria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 88–99, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi:10.18653/v1/2020.coling-main.8.

23 Ahlem Chérifa Khadir, Hassina Aliane, and Ahmed Guessoum. Ontology learning: Grand tour and challenges. *Computer Science Review*, 39, 2021. doi:10.1016/j.cosrev.2020.100339.

24 Javier Lacasta, Javier Nogueras-Iso, and Francisco Javier Zarazaga Soria. *Terminological Ontologies: Design, Management and Practical Applications*, volume 9. Springer Science & Business Media, 2010.

25 Jiao Li, Yueping Sun, Robin Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn Mattingly, Thomas Wiegers, and Zhiyong lu. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, page baw068, 2016. doi:10.1093/database/baw068.

26 Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. arXiv:1907.11692.

27 Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. CamemBERT: a tasty French language model. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online, 2020. Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.645.

28 Roberto Navigli, Paola Velardi, and Juana Maria Ruiz-Martínez. An annotated dataset for extracting definitions and hypernyms from the web. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 2010. European Language Resources Association (ELRA). URL: http://www.lrec-conf.org/proceedings/lrec2010/pdf/20_Paper.pdf.

29 Tapas Nayak and Hwee Tou Ng. Effective modeling of encoder-decoder architecture for joint entity and relation extraction. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8528–8535. AAAI Press, 2020. URL: https://aaai.org/ojs/index.php/AAAI/article/view/6374.

30 Anita Nuopponen. Tangled web of concept relations. concept relations for iso 1087-1 and iso 704. In *Terminology and Knowledge Engineering 2014*, Berlin, Germany, 2014. Association for Computational Linguistics. URL: https://hal.archives-ouvertes.fr/hal-01005882.

31 Giulio Petrucci, Marco Rospocher, and Chiara Ghidini. Expressive ontology learning as neural machine translation. *Journal of Web Semantics*, 52:66–82, 2018. doi:10.1016/j.websem.2018.10.002.

32 Behrang QasemiZadeh and Anne-Kathrin Schumann. The ACL RD-TEC 2.0: A language resource for evaluating term extraction and entity recognition methods. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1862–1868, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL: https://www.aclweb.org/anthology/L16-1294.

33 Bo Qiao, Zhuoyang Zou, Yu Huang, Kui Fang, Xinghui Zhu, and Yiming Chen. A joint model for entity and relation extraction based on bert. *Neural Computing and Applications*, pages 1–11, 2021. doi:10.1007/s00521-021-05815-z.

34 Ayla Rigouts Terryn, Veronique Hoste, Patrick Drouin, and Els Lefever. TermEval 2020: Shared task on automatic term extraction using the annotated corpora for term extraction research (ACTER) dataset. In Béatrice Daille, Kyo Kageura, and Ayla Rigouts Terryn, editors, *Proceedings of the 6th International Workshop on Computational Terminology*, pages 85–94, Marseille, France, May 2020. European Language Resources Association. URL: https://www.aclweb.org/anthology/2020.computerm-1.12.

35 Christophe Roche, Marie Calberg-Challot, Luc Damas, and Philippe Rouard. Ontoterminology: A new paradigm for terminology. In *International Conference on Knowledge Engineering and Ontology Development*, pages 321–326, 2009.

36 Aivaras Rokas, Sigita Rackevičienė, and Andrius Utka. Automatic extraction of lithuanian cybersecurity terms using deep learning approaches. In *Human Language Technologies–The Baltic Perspective*, volume 328, pages 39–46. IOS Press, 2020. doi:10.3233/FAIA200600.

37 Antonio Šajatović, Maja Buljan, Jan Šnajder, and Bojana Dalbelo Bašić. Evaluating automatic term extraction methods on individual documents. In Agata Savary, Carla Parra Escartín, Francis Bond, Jelena Mitrović, and Verginica Barbu Mititelu, editors, *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 149–154, Florence, Italy, 2019. Association for Computational Linguistics. doi:10.18653/v1/W19-5118.

38 Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, undefinedukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc.

39 Lennart Wachowiak, Christian Lang, Barbara Heinisch, and Dagmar Gromann. CogALex-VI shared task: Transrelation - a robust multilingual language model for multilingual relation identification. In Rong Xiang, Emmanuele Chersoni, Luca Iacoponi, and Enrico Santus, editors, *Proceedings of the Workshop on the Cognitive Aspects of the Lexicon*, pages 59–64, Online, 2020. Association for Computational Linguistics. URL: https://www.aclweb.org/anthology/2020.cogalex-1.7.

40 Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771, 2019. arXiv:1910.03771.

41 Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. LUKE: Deep contextualized entity representations with entity-aware self-attention. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454, Online, November 2020. Association for Computational Linguistics. doi:10.18653/v1/2020.emnlp-main.523.

42 Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. DocRED: A large-scale document-level relation extraction dataset. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777, Florence, Italy, July 2019. Association for Computational Linguistics. doi:10.18653/v1/P19-1074.

43 Shuang Zeng, Runxin Xu, Baobao Chang, and Lei Li. Double graph based reasoning for document-level relation extraction. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1630–1640, Online, November 2020. Association for Computational Linguistics. doi:10.18653/v1/2020.emnlp-main.127.

44 Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. Position-aware attention and supervised data improve slot filling. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi:10.18653/v1/D17-1004.

45 Ziqi Zhang, Jose Iria, Christopher Brewster, and Fabio Ciravegna. A comparative evaluation of term recognition algorithms. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May 2008. European Language Resources Association (ELRA). URL: http://www.lrec-conf.org/proceedings/lrec2008/pdf/538_paper.pdf.

46 Suncong Zheng, Yuexing Hao, Dongyuan Lu, Hongyun Bao, Jiaming Xu, Hongwei Hao, and Bo Xu. Joint entity and relation extraction based on a hybrid neural network. *Neurocomputing*, 257:59–66, 2017. doi:10.1016/j.neucom.2016.12.075.

47 Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. Document-level relation extraction with adaptive thresholding and localized context pooling. *CoRR*, abs/2010.11304, 2020. arXiv:2010.11304.