Proceedings of the 6th International Workshop on Computational Terminology (COMPUTERM 2020), pages 62–71
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Towards Automatic Thesaurus Construction and Enrichment

Claudia Lanza†, Amir Hazem§, and Béatrice Daille§
†University of Calabria, Italy
§ LS2N - UMR CNRS 6004, Université de Nantes, France
c.lanza@dimes.unical.it, amir.hazem@ls2n.fr, beatrice.daille@ls2n.fr

Abstract

Thesaurus construction with minimum human efforts often relies on automatic methods to discover terms and their relations. Hence, the quality of a thesaurus heavily depends on the chosen methodologies for: (i) building its content (terminology extraction task) and (ii) designing its structure (semantic similarity task). The performance of the existing methods on automatic thesaurus construction is still less accurate than the handcrafted ones of which is important to highlight the drawbacks to let new strategies build more accurate thesauri models. In this paper, we will provide a systematic analysis of existing methods for both tasks and discuss their feasibility based on an Italian Cybersecurity corpus. In particular, we will provide a detailed analysis on how the semantic relationships network of a thesaurus can be automatically built, and investigate the ways to enrich the terminological scope of a thesaurus by taking into account the information contained in external domain-oriented semantic sets.

Keywords: Automatic thesaurus construction, Terminology extraction, Semantic similarity.

1. Introduction

In computational linguistics and terminology, a thesaurus is often used to represent the knowledge of a specific domain of study as a controlled vocabulary. This paper aims at presenting an analysis of the best performing NLP approaches, i.e., patterns configuration, semantic similarity, morpho-syntactic variation given by term extractors, in enhancing a semantic structure of an existing Italian thesaurus about the technical domain of Cybersecurity. Constructing thesauri by carrying out minimum handcrafted activities is currently highly demanded (Azevedo et al., 2015). Hence, several methods to automatically build and maintain a thesaurus have been proposed so far (Güntzer et al., 1989; Morin and Jacquemin, 1999; Yang and Powers, 2008a; Schandl and Blumauer, 2010). However, the quality of automatically generated thesauri tends to be rather weaker in their content and structure with respect to the conventional handcrafted ones (Ryan, 2014). To guarantee the currency of a thesaurus (Batini et al., 2009) it is crucial to whether improve existing methods or to develop new efficient techniques for discovering terms and their relations. On the perspective of using existing NLP tools for constructing a thesaurus, choosing the most appropriate ones is not an easy task since the performance varies depending on the domain (Nielsen, 2001), the supported languages, the applied strategies, etc. Selecting a highly performing NLP procedure to build on a knowledge representation resource does also contemplate maintenance and enrichment phases aimed at empowering the application usages of these semantic sources.

This work aims at presenting an analysis of which of the NLP approaches, i.e., patterns configuration, semantic similarity, morpho-syntactic variation given by term extractors, could be considered the best performing in enhancing a semantic structure of an existing Italian thesaurus about the technical domain of Cybersecurity. The paper starts firstly from a description of how the current thesaurus has been constructed (Broughton, 2008), following the rules included in the main reference standards for building thesauri (ISO/TC 46/SC 9 2011 and 2013), and of the source corpora composition from which the thesaurus construction has taken its basis. In detail, the paper is organized as follows: Section 2 presents a state-of-the-art on the main works about the construction of terminological knowledge bases, as well as on those that dealt with the semantic relations discovering approaches, such as, the distributional similarity ones. Section 3 describes the former configuration of the handcrafted thesaurus for Cybersecurity and of the source corpus used to build the controlled vocabulary on the Cybersecurity domain, i.e., the Italian corpus made up of legislation and domain-oriented magazines. Section 4 provides an outline of the data sets, i.e., a ranked summary of the terminological lists, including the ones considered as the main gold standards to which rely on; in this part a set of representative examples for each existing relation, which has been extracted from the draft thesaurus to use as data meant to be ameliorated, is given. Section 5 to 7 describe the methods used to automatize the hierarchical, associative and synonymous configuration of the Italian Cybersecurity thesaurus along with their experiments and results. Section 8 combines the results to determine which approach is the best performing with respect to the desired thesaurus output to achieve. Finally, Section 8 presents the conclusion.

2. Objectives

The main purpose presented in this paper is to guarantee a higher-quality management of the Italian Cybersecurity thesaurus' domain-oriented terminology. In particular, this paper explores which could be considered the best performing NLP tool among a plethora of selected ones to be used in order to empower an existing thesaurus of a highly technical domain, the Cybersecurity one. The source language of this semantic resource is the Italian, and the methods pursued to provide reliable candidate terms structures, meant to be included in the thesaurus, are based on sophisticated terminological extractor tools. With the objective of carrying out a study on how to automatically generate the semantic networking systems proper to theauri, these terms extraction software represent the basis from which to be-

gin the non-manually construction of a thesaurus outline. Specifically, the approaches undertaken are the following:

1. Pattern based system: the causative patterns aim at enhancing the associative relationship proper to thesauri configuration;
2. Variants recognition: semantic variation is useful to detect hierarchical and associative sets;
3. Distributional analysis: this procedural methodology helps in identifying the synonymy connection.

Automatically constructing a thesaurus aims at obtaining, as output, an improved knowledge organization system on the Cybersecurity area of study from a semantic correlation construction point of view. This system should provide an advanced hierarchical structuring that is meant to overcome a current thesaurus outline, as well as the associative and equivalence terms organization. Indeed, as described in the following sections, the handcrafted thesaurus categorization sometimes proves to be either subjective and not completely explicit in representing associations among domain-specific terms.

3. Related Works

3.1. Terms Extraction

A thesaurus can be considered as a controlled system that organizes the knowledge of a specific domain of study through a network of semantic relations linked to the hierarchy, synonymy and association structures (Broughton, 2008). Terms included in the thesauri have to keep a unambiguous value, as affirmed in the standard NISO TR-06-2017, Issues in Vocabulary Management: "Controlled vocabulary: A list of terms that have been enumerated explicitly. This list is controlled by and is available from a controlled vocabulary registration authority. All terms in a controlled vocabulary must have an unambiguous, non-redundant definition". Constructing an efficient terminological system usually implies the acquisition of domain-oriented information from texts, specifically those that can provide semantic knowledge density and granularity about the lexicon that is meant to be represented (Barrière, 2006). These structures are in literature known as TKBs (Terminological Knowledge Bases) (Condamines, 2018), and, indeed, they support the modalities of systematizing the specialized knowledge by merging the skills proper to linguistics and knowledge engineering. The ways in which the candidate terms are extracted from a specific domain-oriented corpus (Loginova Clouet et al., 2012) usually follow text pre-processing procedures and extraction of single and multi-word units (Daille and Hazem, 2014) from texts filtered out by frequency measures, then they can undergo a phase of variation recognition (Weller et al., 2011) and other statistical calculations to determine the specificity, accuracy, similarity in the texts from which they come from (Cabré et al., 2001). The reason why the domain-oriented terms are called 'candidates' (Condamines, 2018) is linked to the fact that in the terminologists' activity the need of experts' validation is frequently required, this because just the subjective selection by terminologists might not be exhaustive and fully consistent with the domain expertise (ISO/TC 46/SC 9 2013).

Thesauri's realization is commonly connoted by a manual semantic work that assumes a terminologists' activity in selecting terms from a list of candidate ones, extracted, in turn, from a reference corpus (Condamines, 2007) and, consequently, arranging them in a structure that follows the guidelines given by ISO standards for constructing thesauri (ISO/TC 46/SC 9 2011 and 2013) which aim at normalizing the information meant to be shared by a community of users. For the seek of gaining time to terminologists in defining thesauri's structure (Rennesson et al., 2020), their construction phases are supported by using computer engineering techniques and followed by an evaluation phase that sees experts of the domain involved in the decision-making process about the insertion of the terms in the semantic resource. Even though, a process of appropriateness' check by experts isn't entirely suitable to demonstrate that the TKBs comply with the specialized corpus knowledge flow. Hence, together with certain groups of experts' supervision, other tools support the accuracy validation, i.e., the gold standards (Barrière, 2006). This task is meant to give results on the way terms that have been selected to be part of a semantic resource – designed to represent a specialized language – can be aligned with others included in reference texts. These target texts can be in the same language as the one of the source corpus, and could present less difficulties in the matching system, or multilingual (Terryn et al., 2018), in these cases using translations from existing semantic resources could represent a solution. In this paper, the gold standards taken into account are in Italian language or have been translated in Italian – Nist and Iso – this reflects the native purpose of the project that was intended to provide a guidance for the understanding of the Cybersecurity domain in Italian language.

3.2. Semantic Relations

This paper is going to give a description of the exploited methodologies in automatizing the way thesauri, specifically for the case of study, i.e., Cybersecurity, can be constructed by means of semantic similarity procedures and patterns configuration related to the causative connections. The automatized methodologies used for the configuration of thesauri's structure (Yang and Powers, 2008b; Morin and Jacquemin, 1999), can quicken the process related to the arrangement of textual relations network to shape the informative tissue of a domain. To achieve this framework system different approaches can be pursued, starting from lexico-syntactic patterns conformation (Condamines, 2007), and experimenting other solutions such as the ones proposed by (Grefenstette, 1994) with "Sextant", or (Kageura et al., 2000) with their methodology in considering the common entries in two different thesauri and constructing pairs of codes. As linguistic structures that are very frequent within a corpus of documents (Lefeuvre, 2017), patterns allow to discover among terms which are the conceptual relations (Bernier-Colborne and Barrière, 2018). The study of patterns dates way back, at the end of 90' the works of Hearst (1992) were, for instance, firstly focused on the configuration of Noun Phrases followed by other morpho-syntactic structures to be found

in texts. Many authors in the literature studied the ways nominal and verbal phrases allow to identify semantic relations between terms through syntagmatic or phrasal structures (Girju et al., 2006). The typologies of lexico-syntactic markers help in retrieving the desired semantic information about the terminology proper to a specialized domain (Nguyen et al., 2017), that's the case of the casual relationships between terms. This particular kind of connection is notably described in the works of Barrière (2002) in which the author gives a wide-ranging perspective for investigating the causal relationships in informative texts. As the author underlines, it is not an easy task to group the causative verbs that should isolate the representative terms of a domain to be linked through a cause-effect relation. Nevertheless, grouping some of them can help in identifying the semantic associations to be reflected in a controlled vocabulary given the domain-oriented nature of the casual connections. Indeed, retrieving this type of patterns is a context-dependent procedure: in considering the source area of study and having some technical knowledge about it, terminologists can much easily analyse in an autonomous and accurate way a combination of semantic relationships (Condamines, 2008).

For what concerns semantic similarity methods in the literature, they have firstly been applied to single word terms (SWTs) using a variety of approaches such as: lexicon-based approaches (Blondel and Senellart, 2002), multilingual approaches (Wu and Zhou, 2003; van der Plas and Tiedemann, 2006; Andrade et al., 2013), distributional approaches (Hagiwara, 2008; Hazem and Daille, 2014) and distributed approaches such in (Mikolov et al., 2013; Bojanowski et al., 2016). This procedure helps in configuring the associations between terms with respect to synonyms connections retrieved from corpora. On this point, it is important to highlight the relevance of extracting reliable lists of candidate terms that could represent the starting point from which to set up a conceptual modeling of a thesaurus as well as a basis to analyse and define the internal domain-specific synonyms and hyperonyms (Meyer and Mackintosh, 1996).

4. Thesaurus Structure on Cybersecurity

At this stage, the Italian Cybersecurity thesaurus, on which our paper focuses to describe automatic thesauri construction methodologies, contains 246 terms in the source language (it) and most of them have their definition, or Scope Notes (SN) according to standardized tags (ISO/TC 46/SC 9 2011), taken from the texts from which they derive inside the corpus or the translated gold standards definitions, i.e. Nist and Iso. The thesaurus has been built on the basis of the thesauri construction guidelines from ISO/TC 46/SC 9 2011 and 2013: terms have been formalized in order to guarantee the sharing of information in a standardized way, the concepts of the source corpus have been represented by preferred terms organized according to a network of hierarchical, synonymous and associative semantic relationships. This system allows to set up a knowledge organization oriented towards a creation of semantic connections that, in turn, can create a reflection of the informative scope inside the corpus texts.

The structure phase of the thesaurus for Cybersecurity has started by evaluating the list of terms extracted by using a semi-automatic semantic tool, TextToKnowledge (T2K) (Dell'Orletta et al., 2014), specifically taking into account the frequency scores of the most representative terms and isolating them as being the main candidate terms to be sent to experts' validation process. It was thanks to the co-working process with domain experts that the first list of candidate terms has been filtered out and the first categories, from which the thesaurus structure was developed, provided. This phases resulted after having taken into account several terminological passages:

- the matching process between the output lists derived from the semantic extraction and the taxonomies contained in the gold standards of Nist and Iso; these lists of terms have been translated into Italian language by using an automatic translation software, TRADOS, and a multi/crosslingual terminological platform, IATE;
- the inverse frequency ranks in the term lists;
- the head-term grouping system T2K processed.

In this way, merging the output of a semantic extractor tool, the terminology competencies and the group of experts' validation and supervision, the four main top entry categories have been selected: **Cybersecurity, Cyberbullism, Cyber defence, Cybercriminality**. The goal of the research activity presented in this paper is to improve the decision-making process towards the thesaurus construction by means of approaches that rely on patterns configuration and semantic similarity measures in order to enrich the informative tissue inside the controlled vocabulary.

5. Data Sets

5.1. Corpora

In this section the sets of documents from which the candidate terms have been extracted by using several approaches are presented. The first one refers to the Italian gold standard corpus, i.e., Clusit, and the other, i.e., Cybersecurity corpus, is the one used to build on the source corpus. Taking in consideration a highly specialized field of knowledge with plenty of words in English meant to create a shared base of information among users, the terms extrated resulted to be a hybrid syllabous of English and Italian terms. This because the domain of Cybersecurity owns several technical terms that can be maintained in their English version even providing variants, e.g., *hackers* or *exploit*.

5.1.1. Clusit Corpus

Clusit corpus indicates the reports that have been published by an Italian Cybersecurity organization which shares some of the main cyber threats and attacks together with descriptions, reviews, and a final glossary.

5.1.2. Cybersecurity Corpus

Designing a corpus (Leech, 1991), from which to develop a strong terminological knowledge base that guarantees a rich-context dependency to transmit a reliable representation of a domain, leads to generate a semantic fundamental

dataframe that can be representative of the area of study to be analysed (Condamines, 2018). The Cybersecurity corpus is composed of 220 laws documents and 342 5-sector-oriented magazines. The collection of the texts that compose the source corpus is heterogeneous, this means that the information included takes its ground from legislative documents, regulations, norms, directives, guidelines as well as domain-oriented magazines in order to provide an exhaustive resource to assemble the information representation about the field of knowledge. The information included within the divulgative corpus, with respect to the law data set, provided higher accurate terminology, more targeted kind of concepts to be represented with terms. Table 1 summerizes the number of words (#Words) and the number of documents (#Documents) of the used corpora (Clusit and Cybersecurity).

Corpus	#Words	#Documents
Clusit	385,544	6
Cyber	7,179,829	562

Table 1: Number of words and documents of the Italian corpora: Clusit and Cybersecurity.

5.2. Terminology Lists

For evaluation, we used five terminological lists:

Clusit The Clusit term list contains the main domain specific terms of the reports gathered in a glossary which is composed by a syllabous of these latter followed by their definitions;

Glossary The Glossary term list contains terms with their definitions published by a political intelligence organism, this characteristic has to be taken into account in considering the accuracy and appropriateness of its derived terminology that seems to be weaker than the other more technical domain-oriented resources;

Nist The Nist 7298 - Glossary of Key Information Security Terms (Kisserl, 2013) term list is a complex of terms alphabetically ordered and accompanied by their definitions, also derived from other reference standards. It's considered as a main authoritative data set for Cybersecurity experts on the same level as the Iso list;

Iso The Iso term list refers to the International Standard (ISO/IEC 27000, 2016) for Security and Technology, and it contains, as the Nist, the terms alphabetically ordered with their definitions;

Cyber The Cybersecurity term list contains candidate terms taken from the post-processed texts connected together through the main semantic relationships proper to thesauri (Broughton, 2008), i.e., hierarchical, synonymy, association. These relations are respectively formalized by standard tags (ISO/TC 46/SC 9 2011 and 2013):

broader term broader term (BT) that stands for hyperonyms;

narrower term narrower term (NT) that stands for hyponyms;

used for used for (UF) and **use** (USE) that represent the synonymy relation;

related term related term (RT).

Hereafter some examples of the four addressed relations: hyperonymy (Hyp), synonymy (Syn), related terms (Rel) and cause (Cause).

Hypernym Spam/Phishing, Spam/Smishing, Crypto miner malware/Bitcoin Virus, DoS/DDoS;

Synonym Crackers/Black hat, Sotware malevoli (*malicious software*)/Malware, Cyber minacce (*cyber threats*)/Cyber Threat Actors;

Related Blockchain/Proprietà di sicurezza (*security properties*), Crackers/Hacking, Cyber defence/Cybersecurity;

Causative verb Spoofing/Attacchi informatici (*cyber attacks*) (*to alterate*), Integrità (*integrity*)/Cyber minacce (*cyber threats*) (*to damage*), Attacco (*attack*)/Malware (*implicate*).

Tables 2 and 3 respectively illustrate the size of the term evaluation lists and the distributions of each semantic relation.

	Clusit	Glossary	Nist	Iso	Cyber
#terms	202	284	1282	89	247

Table 2: Size of the 5 term lists.

	Hyp	Syn	Rel	Cause
#terms	172	63	110	68
#pairs	169	35	260	54

Table 3: Semantic similarity evaluation list size. #terms indicates the total number of terms per semantic relation type, and #pairs indicates the number of pairs for each semantic relation.

6. Term Extraction Approaches

6.1. Term Extraction Tools

In this section we provide a description of the chosen tools to execute the terminology extraction.

6.1.1. TermSuite - Variants Detection Tool

TermSuite (Cram and Daille, 2016) is a toolkit for terminology extraction and multilingual term alignment. Its performance is quite immediate when it runs over big data sets. The term extraction provided by TermSuite is a list of representative terms that are presented together with different properties, e.g., their frequency, accuracy, specificity. Terms are therefore ordered according to their unithood and application to the domain. One of the main feature that

shapes the quality of this software is its syntactic and morphological variants detection among terms, e.g, lexical reduction, composition, coordination, derivation (Lanza and Daille, 2019). Variants identification given by the output list in TermSuite represents one of the methods selected to retrieve hyperonyms as well as synonyms in the source corpus. In fact, through the denominative, conceptual and linguistic variants included in the terminological output it is possible to detect in which ways terms are expanded by other semantic elements, reduced, related to an opposite one, or appearing in several linguistic conformations, e.g., *cyber security* or *cybersecurity*. Below a list of few examples to show the variations given by the outputs in TermSuite terminological extraction for Cybersecurity domain in Italian language that can help in detecting semantic associations to be included in the thesaurus:

- **denominative variants:**
  NPN: **hacker (21 matches)** del telefono (*mobile hacker*)→ NA: hacker telefonico
- **conceptual variants:**
  NPN: **worm (8 matches)** → NPNPNA: worm del genere del famigerato nimda (*worm, the infamous nimda kind one*)
- **linguistic variants**:
  N: **antivirus (6 matches)** → A: anti-virus

In the next paragraphs we show how these terms included in the examples above are returned in different ways by the other systems, T2K and PKE.

6.1.2. T2K - Language Design Tool

T2K is an Italian software to automatically extract linguistic information from domain-oriented data sets (Dell'Orletta et al., 2014). The software takes a corpus and processes it according to a default or customized configuration given in input. The list of terms is sorted by the inverse frequency measure or indexed by grouping them according to head-terms ordering. One of the advantages of this semantic extractor is the possibility to personalize the patterns meant to be exploited to execute the extraction of domain-oriented terminology; in this way a more precise semantic chains output can be achieved. On the other hand, though this software shows many benefits related to its flexibility in adapting the configuration to the terminology needs, it performs very slowly when it comes to analyse big corpora. Also for T2K we provide a small set of terms that appear differently from TermSuite's output, or are given with a larger number of results (this is because in T2K the terminological extraction is numerically higher than TermSuite) referred to the aforementioned examples. They provide as well some extra information that can help in orientating the structure outline of the thesaurus blocks:

- **hacker (519 matches)** → hackeraggio (*hacking*)
- **worm (102 matches)** → worm via posta elettronico (*worm via e-mail*)
- **antivirus (127 matches)** → antivirus affetto da trojan (*antivirus affected by trojan*)

6.1.3. Pke - Keyphrases Identification Tool

PKE (Boudin, 2016) is an open-source python keyphrase extraction toolkit that implements several keyphrase extraction approaches. From a linguistic point of view, PKE resulted to be very efficient in terms of providing a semi-automatic structuring of information since many candidate terms, which have been selected as being part of the Cybersecurity thesaurus, are grouped alongside with other ones that, in turn, could represent their associative semantic chains. For this section we provide as well related examples for the terms outputs precision:

- **hacker** and **worm** are found in a same keyphrase cluster → sistemi (*systems*), rete (*network*), worm analisi (*worm analysis*), password, hacker
- **antivirus/anti-virus** not present

New information is on the other hand given by terms that are not appearing in the previous two extractors and that are grouped in a way that can help in structuring their relations inside the thesaurus' outline. In the following cluster it can be observed how the candidate complex term *cyber counterintelligence* could be organized according to the surrounding terms that help in conceiving it as a *technique* or a *procedure* in the cyber intelligence and cyber defence tasks.

> attività (*activities*) intelligence, controspionaggio (*counter espionage*), tecniche (*techniques*), **cyber counterintelligence**, cyber actions, difesa (*defence*)

#cand	T2K	TermSuite	PKE	BERT
Clusit	33,833	15,028	16,664	5,433
CyberSec	593,887	16,541	218,569	6,200

Table 4: Terminology extraction: number of candidate terms extracted by each tool for the Clusit and CyberSec corpora.

6.1.4. BERT

Feature-based approaches are often used for automatic term extraction (Terryn et al., 2018). However, it is often time consuming and not always straightforward to design the most appropriate features to efficiently train a classifier. In order to get rid of the handcrafted features, we chose to apply, as an alternative, a very recent deep neural network approach: Bidirectional Encoder Representations from Transformers (BERT). BERT has proven to be efficient in many downstream NLP tasks (Devlin et al., 2018) including next sentence prediction, question answering, name entity recognition (NER), etc. BERT can be used for feature extraction or for classification. In automatic term extraction (ATE) task, we use BERT as a binary classifier for term prediction. The main idea is to associate each term with its context. Hence, by analogy to next sentence prediction, the first sentence given to BERT is the one which contains the term, and the sentence to predict is the term itself. For training, we feed the model with all the context/term pairs that appear in the corpus as positive examples. The negative examples are generated randomly. Therefore, we hypothesize

		Evaluation lists														
		Clusit			Glossary			Nist			Iso			Cyber		
Corpus	coverage (%)	100			36.2			22.3			55.6			49.3		
	Tools	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
	T2K	0.19	33.1	0.38	0.18	**21.4**	0.36	0.48	**13.1**	0.93	0.09	**37.5**	0.18	0.15	21.4	0.32
Clusit	TermSuite	0.37	27.7	0.73	0.30	15.8	0.58	0.82	9.59	1.51	0.14	23.8	0.27	0.39	23.8	0.77
	PKE	0.85	**69.8**	1.68	0.35	20.4	0.69	0.95	12.4	1.76	0.17	32.9	0.34	0.46	30.7	0.91
	BERT	**2.03**	30.6	**3.81**	**1.03**	16.5	**1.94**	**2.34**	9.59	**3.76**	**0.30**	20.4	**0.59**	**1.07**	**32.7**	**2.07**
	coverage (%)	61.3			72.5			35.3			67.0			100		
	Tools	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
	T2K	0.01	23.7	0.02	0.02	42.2	0.04	0.05	**21.2**	0.10	0.01	**47.7**	0.02	0.01	30.7	0.02
CyberSec	TermSuite	0.10	7.92	0.19	0.37	21.8	0.73	0.77	9.98	1.41	0.12	23.8	0.25	0.20	13.7	0.40
	PKE	0.05	**49.0**	0.10	0.06	**44.0**	0.12	0.12	**21.2**	0.24	0.02	44.3	0.04	0.05	**46.5**	0.10
	BERT	**0.48**	14.3	**0.93**	**1.04**	15.1	**1.95**	**2.30**	7.10	**3.47**	**0.43**	20.4	**0.84**	**1.11**	25.9	**2.13**

Table 5: Terminology extraction results of T2K, TermSuite, PKE and BERT on the Clusit and Cybersecurity corpora. The evaluation is conducted on five lists (Clusit, Glossary, Nist, Iso and Cyber) and the results (%) are given in terms of Precision (P), Recall (R) and F-measure (F1).

that BERT can learn associations between terms and their contexts.

Term extraction systems, with the exception of BERT, include filtering methods that allow the user to set thresholds on various statistical measures above which the ranked candidate terms are kept. In order to favour recall, we decided not to apply any further filtering except for those included as default parameters. Table 4 shows the number of extracted candidates for each used tool/method. T2K software outputs come out with the largest terminological range sets and BERT with the smallest.

6.2. Term Extraction Experiments

We conduct an evaluation on five terminological lists: $Clusit$, $Glossary$, $Nist$, Iso and $Cyber$ and on two corpora: Clusit ($Clusit$) and Cybersecurity ($Cyber$). The results are given in terms of Precision (P), Recall (R) and F-measure (F1). We also give the coverage of each list on each corpus.

Table 5 illustrates the obtained results on the terminology extraction task. Overall, we observe weak results for all the methods. Nonetheless, the recall is much higher especially for PKE and T2K which correlates with the number of their output candidates (see Table 4). The evaluation's list size is very small (around 200) and systems output is often around thousands of terms, which explains the very low precision. Moreover, the evaluation lists are not exhaustive and, by consequence, do not allow a fair evaluation on precision. Indeed, several correct terms which are not present in the evaluation lists have been observed. Finally, based on the F1 score, BERT obtained the best results in all the cases.

7. Semantic Relations Automatization

To address the semantic similarity task, we introduce in the following sections pattern-based and word embedding-based approaches.

7.1. Patterns-based

Among the approaches which have been used for the development of this strategy that could retrieve the semantic connections starting from a domain-oriented data set, the patterns recognition has been one of them (Rösiger et al., 2016). For the purposes of this research activity, some key verbs have been taken into account to represent the causative relationships among the terms included in all the documents of the Italian Cybersecurity source corpus. Almost all of these first verbs imply a relation of agent - cause that provokes some circumstances. The objective of this path-based configuration is to improve the accuracy of the associative relationships included in thesauri and labelled as RT, which stands for *Related Terms* (ISO/TC 46/SC 9, 2011). Indeed, as stated in (Rösiger et al., 2016) work on the achievement of good sets of semantic relationships by employing NLP techniques, the decision of certain verb-object pairs relies on the domain pertinence and relevance, and also on the assumption that these pairs can be syntactically correct. In this step, the verbs considered to launch the queries meant to group the causative relationships among the candidate terms has not followed frequency drills. Almost thirty of the most common casual verbs in Italian have been exploited to retrieve the co-occurrences in the source corpus. The aim about using patterns configuration related to the causative relations (Lefeuvre and Condamines, 2015) is that of providing an improvement in the structure of the related terms in the thesaurus. In ISO Standard 25964 of 2013, when it comes to discuss about the interoperability of the systems, the associative mapping is described as a connection that "[...] may be established between concepts when they do not qualify for equivalence or hierarchical mappings, but are semantically associated to such an extent that documents indexed with the one are likely to be relevant in a search for the other." As can be further observed, the associative relationship in thesauri systematization is among the others, hierarchical and equivalence, the

one that presents more ambiguity in the way it connects the domain-oriented terms. By using causative-based patterns the references from one specific term to another seem more precise and reliable.
The following list presents some examples for the selected causative verbs, some of these relations added new information about the connections among the Cybersecurity specicialized terms, i.e., the relation that occurs between *camouflage* and *password*, or *cyber threats* and the *security properties*; sometimes they confirmed the already configured outline of the thesaurus, as *cyber attacks* and *DDoS* or *spoofing*.

- **provocare (*to provoke*)**:
  virus - worm
  cyber attacks - DDoS
  risks - cyber threats
- **danneggiare (*to damage*)**:
  crackers - data
  cyber threats - integrity, privacy, availability
- **comportare (*to imply*)**:
  cyber attacks - malware
  cyber harrassment - cyber bullism
- **alterare (*to alter*)**:
  camouflage - password
  spoofing - cyber attacks
- **manomesso da (*sabotaged by*)**:
  monitoring - cyber attacks
  monitoring - DoS
- **impattare (*to impact*)**:
  DDoS - cyber attacks
  monitoring - cybersecurity

In summary, causative connections retrieved from source corpus provided added information to the existing ones contained in the Italian Cybersecurity thesaurus, which have already gone through an evaluation phase by a group of experts of the domain.

7.2. Word Embedding-based

Word embedding models have been showing to be very effective in word representation. They have been applied in several NLP tasks including word disambiguation, semantic similarity, bilingual lexicon induction (Mikolov et al., 2013; Arora et al., 2017; Bojanowski et al., 2016), etc. For semantic similarity, and more precisely synonym extraction of multi-word terms, two compositionality-based techniques have been proposed (Hazem and Daille, 2018). The first technique called *Semi-compositional word embeddings* is based on distributional analysis (Hazem and Daille, 2014) and assumes that the head or a tail is shared by two semantically related terms. The second technique called *Full-compositional word embeddings* is inspired by the idea that phrases can be represented by an element-wise sum of the word embeddings of semantically related words of its parts (Arora et al., 2017). In our experiments we follow the principle of the second technique and apply it to the automatic extraction of hyperonyms, synonyms, related and causative terms. The idea is to answer the question: are word embedding models able to extract semantic relations using full-compositionality? All the multi-word terms (MWTs) are represented by a single embedding vector. Each MWT is first characterized by an element-wise sum of its word embedding elements. Then, the cosine similarity measure is applied to extract MWTs synonyms, hypernyms, causative and related terms.

7.3. Semantic Similarity Experiments

We evaluate two word embedding models: word2vec (W2V) (Mikolov et al., 2013) and fastText (Bojanowski et al., 2016). For both models we experiment the Skipgram (Sg) and the Continuous Bag of Words (CBOW) models. The results are shown in terms of precision at 100 (P@100).

	Hyp	Syn	Rel	Cause
W2V (Sg)	**5.91**	**45.7**	5.38	**13.2**
W2V (CBOW)	2.95	34.2	6.15	0.00
fastText (Sg)	4.73	34.2	**10.3**	3.77
fastText (CBOW)	3.55	22.8	**10.3**	1.88

Table 6: Results of semantic relation extraction of word2vec (W2V) and fastText using the Precision at 100 (P@100%) score.

As illustrated in Table 6, all the models fail to extract hypernyms, related, and causative relations. Only synonym extraction exibits acceptable results with Sg (45.7%). Nonetheless, the weak results, even for synonyms can be explained by the mixed nature of language in the Cybersecurity corpus terminology. Indeed, several terms are in English and their related terms in Italian or conversely. This circumstance might weaken the embedding models for low frequency terms.

8. Discussion

To draw guidelines for automatic thesaurus construction, we discuss the following questions: (i) which term extraction system to use; (ii) which system output is the most convenient to enrich an existing term list; (iii) which word embedding model is the most suitable for semantic relation extraction; and, finally, (iv) what kind of relations are extracted by word embedding models. As stated in previous work (Terryn et al., 2018), the evaluation of automatic term extraction is not an easy task. This observation is confirmed in this paper with regards to the obtained results on different evaluation lists (Clusit, Cyber, Iso, etc.). This is particularly true because our evaluation lists are not exhaustive and, for this reason, they don't reflect a real term extraction evaluation scenario. However, they do reflect the situation of thesaurus enrichment, which we stress in this work. If we cannot draw final conclusions on the term extraction performance of the evaluated systems, we can still observe their weak performance on the addressed small subset of terms on Cybersecurity. Nonetheless, this result is to be counterbalanced by encouraging new terms extracted by these systems. Indeed, a manual evaluation of BERT system output, for instance, has shown many new accurate extracted terms. This work represents the first attempt to use BERT model for terminology extraction. Overall, BERT obtained the best results with minimum

supervision and no pattern analysis. This is encouraging since no careful filtering process has been applied, and opens the path for new strategies to pursue for term extraction using deep neural approaches.
For what concerns the types of relations extracted by word embedding models, for the most part the terms in the lists referred to the three semantic relations categories, i.e., hierarchy, association and causative links, prove to be quite similar in the occurrences they provided and, at times, not very faithful, e.g., *cyber gang* is connected in an hierarchical way with *criptography*. On the other hand, the synonyms detection showed better results and the findings are very exhaustive both for what concerns the retrieval of the synonyms themselves, and for the recognition, among the outputs given by the models, of other candidate related terms to add in the thesaurus.
The connections given by these models were performed using the existing thesaurus relations, which have been created following the ISO 25964:2011 rules, as source correspondences to be enhanced with sophisticated grouping procedures. Though the manual evaluation of these series of interrelations has inferred quite similar proximity among the terms extracted in all the four classes of relations, at least on a quantitative level, e.g. *rischi cyber (cyber risks)*, *anti spam*, *hackeraggio (hacking)* appear for almost all the cases, many associated terms helped in improving the thesaural systematization. It should be underlined that when evaluating these kind of lists, a minimum level of knowledge expertise about the technical domain to be studied is required since many terms connected with the head ones sometimes appear related in a very implied way, at least for the domain experts, e.g., *cavalli di troia (trojan horse)* or *zero-day*.
We provide few examples of the additional inputs provided by word embedding techniques on the Italian source corpus about Cybersecurity. It is implied that a new evaluation from the experts of the domain is necessary for the seek of reaching out high pertinence and accuracy levels in the terminological enhanced network meant to transposed in the semantic tool, which is supposed to be shared.

Hyperonyms detection

1. **gestione del rischio cyber (*risk management*)** which has as hyponym *piano di risposta al rischio cyber (risk response measures)*, has been connected with: *attacchi cibernetici (cyber attacks), cavalli di troia (trojan horse), cyber intelligence, difesa informatica (cyber security)*; this confirms the thesaurus outline regarding the top term category of *cybersecurity* and adds another one to be considered, i.e., *cyber intelligence*.

2. **intrusion detection system** - host-based, in the thesaurus is the hyponym of *network security systems*. Among the terms related in a hierarchical way, *network security systems* has been confirmed, and, in turn, other related terms have been included in the semantic structure, e.g., *hacker, mid hacking, sniffing* and *malware*.

Synonymys detection

1. **cybersecurity** has been related to the following synonyms that can be considered as positive candidates for the thesaurus: *difesa informatica (informative defence), deep security, sicurezza cibernetica (cibernetic security), protezione cibernetica (cibernetic protection), sicurezza dei sistemi informativi (informative systems security), sicurezza ict (ict security)*.

2. **malware** has been found related with these synonyms: *software malevolo (malicious software), programmi malevoli (malicious programs)*, confirming the existing synonymous structure in the thesaurus; the interesting result is that *malware* is associated in the same list with several representative terms that will be, in a future perspective, conceived as candidates to improve its semantic connections: *spyware, keylogger, firewall, exploit*.

Related terms detection

1. **zero-day** that in the thesaurus is connected on an associative level with *software vulnerabilities*, is grouped together with *trojan horse, anti spam, hacking, privacy, risk management*.

2. **cyber molestie (*cyber harassment*)**, related in the thesaurus, among others, with *cyber stalking*, has an improved structuring matches since it is found associated also with *cyber theft, hacking, threats, cyber insurance*.

Causative relations detection

1. **cyber minacce (*cyber threats*)** was connected through the causative verb *to damage* to the *security properties of data*, in these models it is linked to *cyber intelligence*, *difesa informatica (cyber security), hackeraggio (hacking) and cavalli di troia (trojan horse)*.

2. **bitcoin** was associated with *data loss* through the causative pattern verb *to prevent*, with the application of these embedding techniques it seems also related with *risk management, cyber risk, spam, hacking*.

9. Conclusion

Automatic thesaurus construction requires efficient methods to collect terminologies and to structure them in a representative way. We discussed in the present paper different approaches for the two building blocks of thesaurus construction: (i) term extraction and (ii) similarity linking. We conducted experiments on an Italian Cybersecurity corpus and reported the performance of existing methods with regards to several evaluation lists. We also proposed a new BERT-based approach that outperformed existing methods on the task of term extraction. If on a general perspective the obtained results provided not so high scores, we observed that system outputs contain accurate candidates that can be used to enrich the existing thesaurus. This is noticeable for the proposed BERT model. Also, regarding semantic similarity, word embedding models showed interesting outputs especially for synonyms and causative relations.

10. References

Andrade, D., Tsuchida, M., Onishi, T., and Ishikawa, K. (2013). Synonym acquisition using bilingual comparable corpora. In *International Joint Conference on Natural Language Processing (IJCNLP'13)*, Nagoya, Japan.

Arora, S., Yingyu, L., and Tengyu, M. (2017). A simple but tough to beat baseline for sentence embeddings. In *Proceedings of the 17th International Conference on Learning Representations (ICLR'17)*, pages 1–11.

Azevedo, C., Iacob, M., Almeida, J., van Sinderen, M., Ferreira Pires, L., and Guizzardi, G. (2015). Modeling resources and capabilities in enterprise architecture: A well-founded ontology-based proposal for archimate. *Information systems*, 54:235–262, 12.

Barrière, C. (2002a). Hierarchical refinement and representation of the causal relation. *Terminology*, 8(1):91–111.

Barrière, C. (2002b). Investigating the causal relation in informative texts. *Terminology*, 7(4):135–154.

Barrière, C. (2006). Semi-automatic corpus construction from informative texts. In Lynne Bowkes, editor, *Text-Based Studies in honour of Ingrid Meyer*, Lexicography, Terminology and Translation, chapter 5. University of Ottawa Press, January.

Batini, C., Cappiello, C., Francalanci, C., and Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Comput. Surv.*, 41(3), July.

Bernier-Colborne, G. and Barrière, C. (2018). CRIM at semeval-2018 task 9: A hybrid approach to hypernym discovery. In *Proceedings of The 12th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT, New Orleans, Louisiana, June 5-6, 2018*, pages 725–731.

Blondel, V. D. and Senellart, P. (2002). Automatic extraction of synonyms in a dictionary. In *SIAM Workshop on Text Mining*.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016). Enriching word vectors with subword information. *CoRR*, abs/1607.04606.

Boudin, F. (2016). pke: an open source python-based keyphrase extraction toolkit. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 69–73, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Broughton, W. (2008). *Costruire Thesauri: strumenti per indicizzazione e metadati semantici"*. EditriceBibliografica, 2008, Milano, Italia Cliffs, NJ.

Cabré, M. T., Bagot, R. E., and Platresi, J. V. (2001). Automatic term detection: A review of current systems. In *Recent Advances in Computational Terminology*, volume 2 of *Natural Language Processing*, pages 53–88. John Benjamins.

Condamines, A. (2007). L'interprétation en sémantique de corpus : le cas de la construction de terminologies. *Revue française de linguistique appliquée*, Vol. XII(2007/1):39–52.

Condamines, A. (2008). Taking Genre into account when Analyzing Conceptual Relation Patterns. *Corpora*, 8:115–140.

Condamines, A. (2018). Terminological knowledge bases from texts to terms, from terms to texts. In *The Routledge Handbook of Lexicography*. Routledge.

Cram, D. and Daille, B. (2016). Terminology extraction with term variant detection. In *Proceedings of ACL-2016 System Demonstrations*, pages 13–18, Berlin, Germany, August. Association for Computational Linguistics.

Daille, B. and Hazem, A. (2014). Semi-compositional method for synonym extraction of multi-word terms. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1202–1207, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Dell'Orletta, F., Venturi, G., Cimino, A., and Montemagni, S. (2014). T2K: a system for automatically extracting and organizing knowledge from texts. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Girju, R., Badulescu, A., and Moldovan, D. (2006). Automatic discovery of part-whole relations. *Computational Linguistics*, 32(1):83–135.

Grefenstette, G. (1994). *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publisher, Boston, MA.

Güntzer, U., Jüttner, G., Seegmüller, G., and Sarre, F. (1989). Automatic thesaurus construction by machine learning from retrieval sessions. *Inf. Process. Manage.*, 25(3):265–273, May.

Hagiwara, M. (2008). A supervised learning approach to automatic synonym identification based on distributional features. In *Proceedings of the ACL-08: HLT Student Research Workshop*, pages 1–6, Columbus, Ohio, June. Association for Computational Linguistics.

Hazem, A. and Daille, B. (2014). Semi-compositional method for synonym extraction of multi-word terms. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Hazem, A. and Daille, B. (2018). Word Embedding Approach for Synonym Extraction of Multi-Word Terms. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 7-12, 2018. European Language Resources Association (ELRA).

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *COLING 1992 Volume 2: The 15th International Conference on Computational Linguistics*.

ISO/IEC 27000, (2016). *Information technology – Security techniques – Information security management sys-*

tems – Overview and vocabulary. International Standard, February.

ISO/TC 46/SC 9, (2011). *ISO 25964-1:2011 Information and documentation — Thesauri and interoperability with other vocabularies — Part 1: Thesauri for information retrieval*. International Standard, August.

ISO/TC 46/SC 9, (2013). *ISO 25964-2:2013 Information and documentation — Thesauri and interoperability with other vocabularies — Part 2: Interoperability with other vocabularies*. International Standard, March.

Kageura, K., Tsuji, K., and Aizawa, A. N. (2000). Automatic thesaurus generation through multiple filtering. In *Proceedings of the 18th Conference on Computational Linguistics - Volume 1*, COLING '00, page 397–403, USA. Association for Computational Linguistics.

Kisserl, R., (2013). *Glossary of Key Information Security Terms*. National Institute of Standards and Technology, May. NISTIR 7298 Revision 2.

Lanza, C. and Daille, B. (2019). Terminology systematization for Cybersecurity domain in Italian language. In *TIA 2019 Terminologie et Intelligence Artificielle - Atelier TALN-RECITAL et IC (PFIA 2019)*, Toulouse, France, July.

Leech, G. (1991). *The state of the art in corpus linguistics*. Longman, London.

Lefeuvre, L. and Condamines, A. (2015). Constitution d'une base bilingue de marqueurs de relations conceptuelles pour l'élaboration de ressources terminoontologiques. In *Terminology and Artificial Intelligence (TIA'2015)*, pages 183–190, Granada, Spain.

Lefeuvre, L. (2017). *Analyse des marqueurs de relations conceptuelles en corpus spécialisé : recensement, évaluation et caractérisation en fonction du domaine et du genre textuel*. Ph.D. thesis. Thèse de doctorat Sciences du langage - U. Toulouse 2.

Loginova Clouet, E., Gojun, A., Blancafort, H., Guegan, M., Gornostay, T., and Heid, U. (2012). Reference Lists for the Evaluation of Term Extraction Tools. In *Terminology and Knowledge Engineering Conference (TKE)*, Madrid, Spain.

Meyer, I. and Mackintosh, K. (1996). The Corpus from a Terminographer's Viewpoint. *International Journal of Corpus Linguistics*, 1(2):257–285.

Mikolov, T., Yih, S. W.-t., and Zweig, G. (2013). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT-2013)*. Association for Computational Linguistics, May.

Morin, E. and Jacquemin, C. (1999). Projecting corpus-based semantic links on a thesaurus. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, page 389–396. Association for Computational Linguistics.

Nguyen, K. A., Köper, M., Schulte im Walde, S., and Vu, N. T. (2017). Hierarchical embeddings for hypernymy detection and directionality. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 233–243. Association for Computational Linguistics.

Nielsen, M. L. (2001). A framework for work task based thesaurus design. *Journal of Documentation*, 57(6):774–797.

Rennesson, M., Georget, M., Paillard, C., Perrin, O., Pigeotte, H., and Tête, C. (2020). Le thésaurus, un vocabulaire contrôlé pour parler le même langage. *Médecine Palliative*, 19(1):15 – 23. Documentation et pratiques documentaires en soins palliatifsCoordonné par Caroline Tête.

Ryan, C. (2014). Thesaurus construction guidelines: An introduction to thesauri and guidelines on their construction. *Dublin: Royal Irish Academy and National Library of Ireland*.

Rösiger, I., Bettinger, J., Schäfer, J., Dorna, M., and Heid, U. (2016). Acquisition of semantic relations between terms: how far can we get with standard nlp tools? In *Proceedings of COLING 2016: 5th International Workshop on Computational Terminology (CompuTerm)*, Osaka, Japan.

Schandl, T. and Blumauer, A. (2010). Poolparty: Skos thesaurus management utilizing linked data. In *The Semantic Web: Research and Applications*, pages 421–425, Berlin, Heidelberg. Springer Berlin Heidelberg.

Terryn, A. R., Hoste, V., and Lefever, E. (2018). A Gold Standard for Multilingual Automatic Term Extraction from Comparable Corpora: Term Structure and Translation Equivalents. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 7-12, 2018. European Language Resources Association (ELRA).

van der Plas, L. and Tiedemann, J. (2006). Finding synonyms using automatic word alignment and measures of distributional similarity. In *21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics ACL'06*, Sydney, Australia.

Weller, M., Gojun, A., Heid, U., Daille, B., and Harastani, R. (2011). Simple methods for dealing with term variation and term alignment. In *Proceedings of the 9th International Conference on Terminology and Artificial Intelligence (TIA 2011)*, pages 87–93, Paris, France, November. INALCO.

Wu, H. and Zhou, M. (2003). Optimizing synonym extraction using monolingual and bilingual resources. In *Proceedings of the second international workshop on Paraphrasing-Volume 16*, pages 72–79. Association for Computational Linguistics.

Yang, D. and Powers, D. M. (2008a). Automatic thesaurus construction. In *Proceedings of the Thirty-First Australasian Conference on Computer Science - Volume 74*, ACSC '08, page 147–156, AUS. Australian Computer Society, Inc.

Yang, D. and Powers, D. M. (2008b). Automatic thesaurus construction. In *Proceedings of the Thirty-first Australasian Conference on Computer Science - Volume 74*, ACSC '08, pages 147–156, Darlinghurst, Australia, Australia. Australian Computer Society, Inc.