






Knowledge Graph Generation from Text Using Supervised Approach Supported by a Relation Metamodel: An Application in C2 Domain

Jones O. Avelino^{1,2}, Giselle F. Rosa¹, Gustavo R. Danon¹, Kelli F. Cordeiro³
and Maria Cláudia Cavalcanti¹

¹*Instituto Militar de Engenharia (IME), Rio de Janeiro, RJ, Brazil*

²*Centro de Análise de Sistemas Navais (CASNAV), Rio de Janeiro, RJ, Brazil*

³*Subchefia de Comando e Controle (SC-1), Ministério da Defesa, Brasília, DF, Brazil*

Keywords: Named Entity Recognition, Relation Extraction, Knowledge Graph, Command and Control.

Abstract: In the military domain of Command and Control (C2), doctrines contain information about fundamental concepts, rules, and guidelines for the employment of resources in operations. One alternative to speed up personnel (workforce) preparation is to structure the information of doctrines as knowledge graphs (KG). However, the scarcity of corpora and the lack of language models (LM) trained in the C2 domain, especially in Portuguese, make it challenging to structure information in this domain. This article proposes IDEA-C2, a supervised approach for KG generation supported by a metamodel that abstracts the entities and relations expressed in C2 doctrines. It includes a pre-annotation task that applies rules to the doctrines to enhance LM training. The IDEA-C2 experiments showed promising results in training NER and RE tasks, achieving over 80% precision and 98% recall, from a C2 corpus. Finally, it shows the feasibility of exploring C2 doctrinal concepts through an RDF graph, as a way of improving the preparation of military personnel and reducing the doctrinal learning curve.


1 INTRODUCTION


Military performance in the Command and Control (C2) scenario may be impacted by personnel turnover, which is inherent to military careers. Thus, the Armed Forces (AF) provide a list of doctrinal documents comprising a set of principles, concepts, standards, and procedures that guide actions and activities for the full employment of its personnel in military operations and exercises. Despite this, studying these documents can lead to a long and costly learning curve. On the other hand, as educational sources, they serve for extracting helpful and structured information, which could shorten the learning curve (Chaudhri et al., 2013).


Advances in the Information Extraction (IE) technique in Natural Language Processing (NLP) have made it possible to extract data from texts (structured,


semi-structured, and unstructured) through Named Entity Recognition (NER) and Relation Extraction (RE), based on the search for occurrences of object classes (Luan et al., 2018). Since the emergence of the self-attention mechanism and Language Models (LM) based on Transformers, it has been possible to expand NLP tasks (Devlin et al., 2019). By training an LM with examples from the domain, it is possible to create a specialized LM (Lee et al., 2019). On the other hand, approaches that train LMs with fixed categories of entities limit their application, the extraction of knowledge, and the expansion of the trained LM.


This work aims to minimize this limitation using the IDEA-C2 approach, a supervised approach that supports the generation of KG based on the training of LM from C2 doctrinal texts in Portuguese. To support the training, the approach encompasses pre-annotation and curation processes, both supported by a metamodel that defines high-level constructs to annotate the texts. In addition, the metamodel supports the generation of the KG based on the mapping of its constructs to the resources of controlled vocabularies or the approach itself. To this end, we implemented the IDEA-C2-Tool prototype, which uses the

^a  <https://orcid.org/0000-0001-9483-7220>

^b  <https://orcid.org/0009-0004-8512-7883>

^c  <https://orcid.org/0009-0005-2881-6030>

^d  <https://orcid.org/0000-0001-5161-8810>

^e  <https://orcid.org/0000-0003-4965-9941>

BERTimbau LM to perform the training. By submitting C2 Texts to the trained LM, it extracts by inference sets of entities and relations, which can then be explored through an RDF graph. The contributions of this work include: (i) an approach to support extracting entities and relations and generating knowledge graphs; (ii) a prototype that implements the activities of the approach; and (iii) an experiment that demonstrates the viability and usefulness of IDEA-C2.

2 BACKGROUND

Machine Learning algorithms have been used to train LM by different approaches. The supervised approach is characterized by the work of the domain expert, in addition to the need for a corpus of annotated texts based on categories of entities and relations (Russell and Norvig, 2010). The text annotation task is the identification of which category is appropriate for a given term. In NER tasks, named entities are categorized, for example, as person, organization, and location, while in RE tasks, the categories are used to express the semantics between two named entities, such as born-in, married-to, etc. However, manually annotating the corpus is very time-consuming. The supervised distance approach emerged as an alternative to minimize annotation costs (Mintz et al., 2009). It uses regular expression rules to automate the annotation task.

The Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) is an LM that allows for training models based on examples from the domain. BERT training consists of two stages. The first stage is pre-training, feature-based, and does not require labeled data. The second stage involves fine-tuning the weights of the pre-trained model to adjust it based on the domain dataset (Devlin et al., 2019). At this stage, the categories of entities and relations are usually defined with the help of the domain expert and according to the application domain. This is not an easy task and it has been recognized as so by database and conceptual modelers for decades (Kent, 2012).

Due to the difficulty of identifying categories of entities and relations for any domain, a viable alternative is to use a metamodel that allows abstracting and flexibilizing this definition using high-level constructs. To represent constructs of different abstraction levels (models and metamodels) in a single view, it is necessary to use flexible modeling approaches, such as Knowledge Graphs (KG). As in (Hogan et al., 2021), a KG is a graph of (meta)data intended to accumulate and convey knowledge of the real world,

whose nodes represent entities of interest and whose edges represent relations between these entities. The Resource Description Framework (RDF)¹ is a largely used implementation of KG. It represents (meta)data as a directed graph, made up of triples, formed by a subject, a predicate, and an object (s, p, o), where *subjects* and *objects* correspond to the vertices of the graph, and *predicates* correspond to the edges.

3 RELATED WORK

In general, works focused on generating KG applying LM are diverse. However, they share some common characteristics. One of these is the use of relation extraction to create triples. In (Liu et al., 2023), an aviation field KG is generated from textbook chapter texts. Pairs of entities and relations are extracted and combined with reinforcement learning methods, using five entity categories and three relations. The Hidden Markov Model (HMM), Conditional Random Field model (CRF), Bidirectional Long Short-Term Memory (BiLSTM), and BiLSTM + CRF are used for this purpose. However, the definition of these categories limits training. In addition, the Transformers architecture outperforms these models by searching for more distant terms in a bidirectional manner.

In (Dang et al., 2023), a KG is created based on extractions of five categories of entities and relations from nutrition and mental health PubMed articles. A hybrid model deals with NER tasks, supported by ontologies. For RE, the authors applied a model that combines patterns of word syntactic dependencies with part of speech in a sentence. To this end, scispaCy², a pipeline of models based on biomedical data, is used. However, the approach is limited to fixed categories. In addition, using supervised distance methods outperforms syntactic dependency methods (Mintz et al., 2009).

In (Zhou et al., 2022), the supervised distance method is used to minimize manual annotation. Military simulation scenarios are established based on NER tasks. Four categories of entities are defined for training using a recurrent neural network with short- and long-term memory (LSTM). The LSTM learns the dependencies between elements in a sequence. An Embedding layer converts text into a vector representation. Another BiLSTM layer, made up of two LSTMs in opposite directions, extracts the context. Finally, the entities are converted into class diagrams to transform them into RDF graphs. Despite minimizing annotation, the categories of entities are fixed.

¹<https://www.w3.org/RDF/>

²<https://allenai.github.io/scispaCy/>

In addition to the fact that the approach does not deal with RE tasks, it impacts analysis in RDF.

Finally, in (Zhao et al., 2021), a KG based on military regulations is generated. The annotation is supported by a statistical word segmentation method combined with dependency parsing for NER tasks. For the RE task, the authors applied Conditional Random Fields and Part-of-speech tagging (POS) to indicate the grammatical class of the word that denotes the action between the pairs of entities, obtaining the triples e_i, r_j, e_k for generating the KG. It should be noted that recognizing entities without defining fixed categories is an aspect to be considered. However, the extraction of relations based on POS is limited to the structure of the text.

Table 1 presents the related works and the comparison parameters to our proposal. We can highlight three important characteristics to evaluate in these works: (i) approaches to minimize the impact of manual annotation, which contribute to a greater number of labeled data; (ii) good recall of the domain, i.e., not limited to a fixed set of entity categories; (iii) usage of didactic texts, glossaries or ontologies to increase the LM adherence to the domain.

Table 1: Comparison with related works.

Approach	<i>i</i>	<i>ii</i>	<i>iii</i>
(Liu et al., 2023)	X	-	X
(Dang et al., 2023)	-	-	X
(Zhou et al., 2022)	X	-	X
(Zhao et al., 2021)	-	X	-

Although the approaches apply various strategies to generate KGs, the problem of defining annotation categories remains open. One of the challenges is to find a representation that can deal with more flexible categories of entities and relations. Unlike these other approaches, this work aims to offer a solution that meets the three characteristics.

4 C2RM: COMMAND AND CONTROL RELATIONS METAMODEL

The proposed Command and Control Relations Metamodel (C2RM) aims to define a structure to represent the recognized entities and provide the semantics of the relations between them based on the use of doctrinal texts and glossaries of terms of the C2 domain. To address the challenge of dealing with the flexibility of categories, the C2RM defines high-level abstraction constructs, **Entity** and **Relation**, capable of providing

comprehensive categories for extracting information from the corpus, including relations of general application (such as term-definition, hyperonym-hyponym, whole-part, equivalent-synonym) as well as C2 domain relations (such as action-responsible).

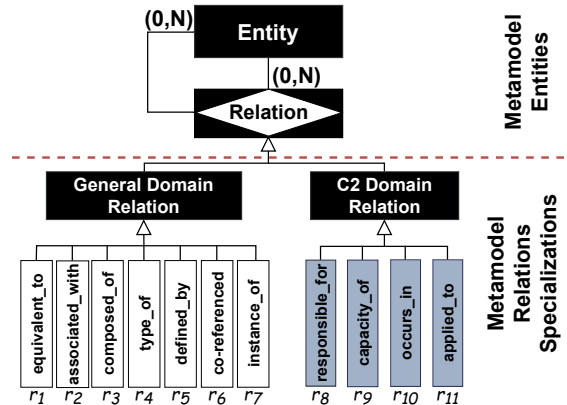


Figure 1: C2RM Diagram.

As illustrated in Figure 1, the C2RM represents two high-level constructs: **Entity** and **Relation**. The **Entity** construct, $E = \{e_1, e_2, \dots, e_n\}$, represents the named entities recognized from the text, for example, Person, Brigade, Operation, Alpha Operation, etc. Similarly, the **Relation** construct, $R = \{r_1, r_2, \dots, r_m\}$, represents the instances of relations that may occur between two Entity instances. Note that it was a choice not to represent a predefined set of entity categories. There is just a single generic category, the ENTITY. The idea is to increase the flexibility of the approach, named Singlecategory classification. On the other hand, the **Relation** construct was specialized into Multicategory classifications. It has two specializations: **General Domain Relation** and **C2 Domain Relation**, which represent the relations outside and inside the C2 domain, respectively. The self-relationship aggregation has characteristics that allow each relationship to be specialized. Also, it has eleven sub-specializations³ which are defined to represent the semantics of the relationships. Specializations r_1, r_2 , and r_7 were inspired by RDF properties, denoting equivalence, association, and instance, respectively. Specializations r_3 and r_4 were inspired by (Augenstein et al., 2017), denoting compositions and hierarchies, while r_5 and r_6 were inspired by (Spala et al., 2020), denoting term-definition and co-reference, respectively. Finally, r_8 to r_{11} are specializations involving C2 domain, denoting responsibility, capacity, occurrence, and application, respectively.

The main benefit of the C2RM is that it allows

³Although the specializations are in English, they express semantics in Portuguese

one to work only with pre-established relations, most of them general-domain relations, and some of them C2-related relations, but that can also be considered generic to a certain extent. Besides, all these relations may be identified in texts at multiple levels of abstraction. Sometimes they appear at the instance level, and sometimes they may be seen as connecting high-level concepts. In the sentence “Operação de Garantia da Lei e da Ordem - Operação Militar conduzida pelas Forças Armadas...”, extracted from a C2 Glossary (BRASIL, 2009), it was annotated that “Operação de Garantia da Lei e da Ordem” and “Forças Armadas” are instances of the **Entity** construct, and are connected by an instance of the **Relation** construct *responsible_for*. In the sentence “Fica autorizado o emprego das Forças Armadas (Marinha do Brasil,...) para a Garantia da Lei e da Ordem”, extracted from the Presidential Decree establishing the military operation to Guarantee Law and Order (GLO) in 2017⁴, it was annotated that “Garantia da Lei e da Ordem (GLO 2017)” and “Marinha do Brasil” are instances of the **Entity** construct, and are connected by an instance of the **Relation** construct *responsible_for*. Note that at this point, there is no information about the categories of those **Entity** instances. However, an additional annotation of instances of the **Relation** construct *instance_of*, connects “Marinha do Brasil” to “Forças Armadas”, and connects “GLO 2017” to “Operação de Garantia da Lei e da Ordem”. This example illustrates that with C2RM metamodel it is possible to generate a domain model (C2 Model) with two levels of abstraction. From the first sentence two high-level concepts (categories) are identified, and from the second sentence two instances of those concepts are identified, both pairs are connected through an instance of the **Relation** construct *responsible_for*.

5 IDEA-C2: KG GENERATION FROM TEXT SUPPORTED BY C2RM

The IDEA-C2 (generation of knowledge graphs based on Artificial intelligence of C2 Domain) supervised approach is a process made up of seven sub-processes, illustrated in Figure 2. The IDEA-C2 aims to generate KG in the C2 domain, in Portuguese, supported by the BERTimbau LM (Souza et al., 2020), from a corpus of semi-structured texts, which are based on C2 glossaries and doctrinal documents. In

⁴https://www.planalto.gov.br/ccivil.03/_ato2015-2018/2017/dsn/dsn14485.htm

addition, IDEA-C2 uses the C2RM that contributes to the pre-annotation of the input texts, the curation, the fine-tuning of the LM and the generation of the KG.

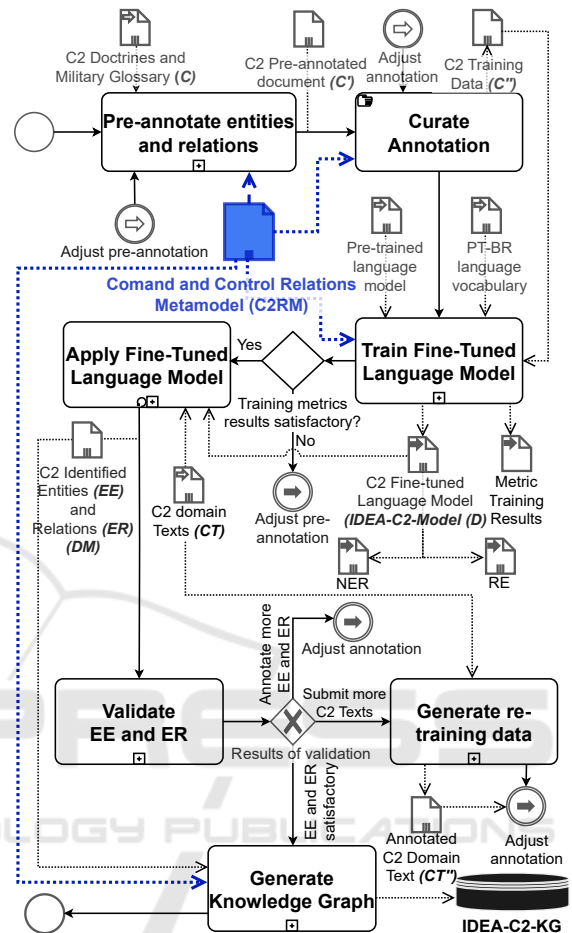


Figure 2: Overview of IDEA-C2 Approach.

Departing from a set of doctrine texts named *UC2*, which constitutes a C2 Corpus, a representative subset of it (*C Corpus*) is selected and submitted to the IDEA-C2 approach. The *C Corpus* is first annotated using the C2RM constructs, and then submitted to BERTimbau LM for NER and RE training, resulting in the IDEA-C2-Model. In reality, IDEA-C2-Model represents two trained LMs, one for the NER task and the other for RE. Another sample of C2 Texts is then submitted to IDEA-C2-Model in order to extract named entities and their relations. The extracted data is stored in the KG Database, generating IDEA-C2-KG. Next, the IDEA-C2 sub-processes are presented in detail.

The Pre-annotate entities and relations sub-process has as its input the sentences, s_i , from the unlabeled doctrines and military glossary text corpus, represented by $C = \{s_1, s_2, \dots, s_n\}$, for pre-

annotation. Pre-annotation was inspired by the distance supervision method (Mintz et al., 2009) in order to increase the labeling of terms and minimize the curation effort. Using the specializations of the C2RM, detailed in Section 4, pre-annotation rules are developed using regular expressions to annotate terms, generating as output a new pre-annotated corpus, C' , in JSON Lines (JSONL) format.

The Curate Annotation sub-process has as its input the corpus, C' , for the expert to curate. The curator in the supervised approach can either revise or ratify the annotated entities and relations or insert new annotations. The Doccano tool⁵, previously configured with the C2RM constructs, supports the curator. At the end, a new corpus is generated, C'' , C2 training data, containing the finalized annotations.

The Train Fine-Tuned Language Model sub-process has as its input the corpus, C'' , the BERTimbau LM, the Portuguese language vocabulary and the C2RM. It also submits C'' , annotated with the categories from the C2RM, to BERTimbau LM for training in order to identify named entities and extract relations. Initially, the sentences of C'' are retrieved, standardized to lowercase and stopwords removed. In the tokenization activity, the SpaCy⁶ library pipeline is used, which retrieves each sentence s_i , splits their terms into tokens and transforms these tokens into identifiers to create a spaCy **Doc** object.

The D dataset is split into training, validation, and test sets, that are used to train the IDEA-C2-Model, the NER/RE model for the C2 domain. After that, the precision and recall metrics are evaluated, as well as the inferences from the NER and RE tasks identified by IDEA-C2-Model. If the results are not satisfactory, it returns to Pre-annotate entities and relations for to pre-annotate. Otherwise, once the IDEA-C2-Model is ready, the Apply Fine-Tune language Model sub-process is activated, and may submit a subset of not annotated C2 texts ($CT \subset UC2 - C$) to the NER and RE tasks. Thus, this sub-process has as its input $CT = \{st_1, st_2, \dots, st_m\}$, and as the output the DM dataset, consisting of entities, $EE = \{e_1, e_2, \dots, e_n\}$, and the triples of relations, $ER = \{(e_i, r_j, e_k) \mid e_i, e_k \in EE \text{ are semantically related by } r_j \in R, \text{ in some } st_l \in CT\}$.

In the Validate EE and ER sub-process, the curator is responsible for evaluating the inferences of the NER and RE tasks identified by IDEA-C2-Model through EE and ER compatible with the named entities and CT relations. If the results are satisfactory, the Generate Knowledge Graph sub-process is activated. Otherwise, the curator can either choose to re-evaluate the annotation, in which case it returns

to Cure pre-annotation, or the curator can introduce more texts into the C2 domain, in which case Generate re-training data is activated.

In the Generate re-training data sub-process, the curator can reinput the CT texts that were submitted to IDEA-C2-Model. In this case, IDEA-C2 retrieves CT , including the annotations already identified by IDEA-C2-Model, has as its output the corpus CT'' as a result. The sub-process Curate Annotation is activated with the CT'' texts for the curator to review and/or include new annotations of both named entities and relations.

Finally, after the iterative cycles of curation and retraining, the Generate Knowledge Graph sub-process retrieves the entities, EE , the triples, ER , and the properties, R , in order to generate IDEA-C2-KG. Initially, the entities, EE , are created as `rdfs:Class`. The triples ER are created according to the mapping, R , between the specializations of the metamodel and the properties of the RDF graph, as expressed in Table 2. In addition, the namespace `c2rm` was created to deal with specializations with no corresponding property in the RDF graph, as in the following cases: r_6 , r_8 , r_9 , r_{10} and r_{11} .

Table 2: Mapping between C2RM and the RDF Graph.

r_n	Specializations of C2RM	RDF Property
r_1	equivalent_to	owl:equivalentClass
r_2	associated_with	rdfs:seeAlso
r_3	composed_of	rdf:Bag
r_4	type_of	rdfs:subClassOf
r_5	defined_by	rdfs:comment
r_6	co-referenced	c2rm:coreferenced
r_7	instance_of	rdf:type
r_8	responsible_for	c2rm:responsible_for
r_9	capacity_of	c2rm:capacity_of
r_{10}	occurs_in	c2rm:occurs_in
r_{11}	applied_to	c2rm:applied_to

An example of a C2 Text submitted to the IDEA-C2-Model is described as follows. Previously, the Pre-annotate Entities and Relations sub-process, using distance-supervised methods, annotated relations e_1, e_3, e_4, e_7 and e_8 of Table 3, while the others were manually annotated at the Curate Annotation process. In addition, the following relations were also manually annotated: `type_of`, `capacity_of`, and `applied_to`. These annotations were used as input to the Train Fine-Tuned Language Model sub-process, generating the IDEA-C2-Model. In the example, the following sentence st_1 of CT was submitted to the sub-process Apply Fine-Tuned Language Model: “*Os elementos do poder de combate terrestre representam*

⁵<https://github.com/doccano/doccano>.

⁶<https://spacy.io/>

a essência das capacidades que a F Ter emprega em situações – sejam de guerra ou de não guerra. São eles: Liderança, Informações e as Funções de Combate.”⁷. The result was that all the annotated entity and relation instances (Table 3) were identified by the application of the IDEA-C2-Model.

Table 3: Entities and Relations identified by the IDEA-C2.

(e_n)	Entities	Relations
e_1	Elementos do poder de combate	-
e_2	elementos do poder de combate terrestre	(e_2, r_4, e_1)
		(e_2, r_9, e_3)
		(e_2, r_{11}, e_4) (e_2, r_{11}, e_5)
e_3	F Ter	-
e_4	guerra	-
e_5	não guerra	-
e_6	Liderança	(e_6, r_4, e_2)
e_7	Informações	(e_7, r_4, e_2)
e_8	Funções de Combate	(e_8, r_4, e_2)

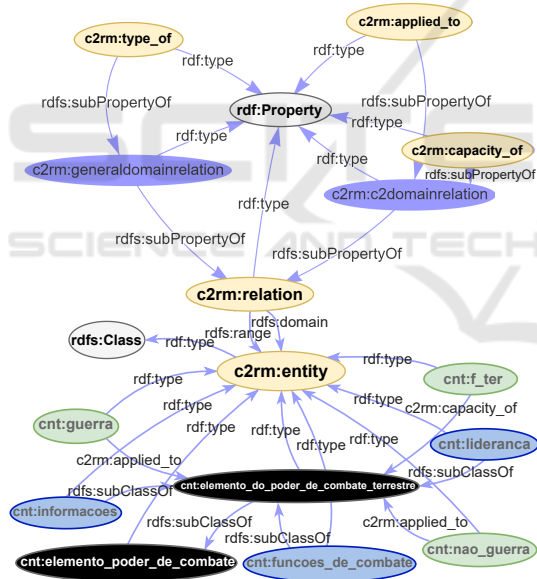


Figure 3: Example of IDEA-C2-KG: RDF graph result.

In Figure 3, the resources in gray, rdfs:Class and rdf:Property, are RDF metaclasses. In addition, in yellow, c2rm:entity, relation, type_of, capacity_of and applied_to represent the constructs of the C2RM. Moreover, cnt is the namespace of IDEA-C2-KG. The resources in green, cnt:nao.guerra,

guerra and f_ter are entities. As the specialization “type_of” is a relation of hyperonymy and hyponymy between two entities, the superclasses are represented by cnt:element_do_poder_de_combate and poder_de_combate_terrestre in black. The subclasses, in blue, are represented by cnt:lideranca, funcoes_of_combate and informacoes and are related to the superclass through the property rdfs:subClassOf.

6 EXPERIMENTS AND RESULTS

To validate the IDEA-C2 approach with C2RM support, two experiments were performed. They showed promising results in terms of flexibility in the annotation of entities and relations and of the training subprocess performance. To carry out the experiments the IDEA-C2-Tool⁸ was developed in Python v.3 using the spaCy pipeline with Transformers component, spacy-transformers.TransformerModel.v3⁹.

6.1 Annotation Strategy Based on Singlecategory NER Classification

The first experiment aimed to validate the strategy for defining high-level C2RM constructs in the IDEA-C2 approach. To this end, the LM was trained for the NER and RE tasks by submitting two corpora, the SciERC with 500 scientific abstracts annotated with scientific entities, their relations and coreference clusters (Luan et al., 2018) and Material Science with 800 abstracts annotated manually (Weston et al., 2019). After training the LM, the results of the training metrics were collected. Table 4 shows the experiment results of the IDEA-C2 approach, which is based on a single category strategy for annotating entities (**Singlecategory NER**), and on the multicategory for annotating relations **Multicategory RE**, compared to the results of the application of the multicategory strategy for both NER and RE tasks.

The Train Fine-Tuned Language Model subprocess, implemented as a spaCy pipeline, was configured as follows. For both corpora, the Dropout was set to 20%, as the tests showed good results. Similarly, the vocabulary used was en_core_web_sm because the corpora were in English. Finally, the BERT Model used as input (highlighted in Table 4), was set to two different values. In the case of SciERC corpus, we set it to allenai/scibert LM, according to (Luan et al., 2018). For the Material Science corpus, we

⁷Translation: The elements of ground combat power represent the essence of the capabilities that the F Ter employs in situations - both war and non-war. They are: Leadership, Intelligence and the Combat Functions.

⁸<https://github.com/jonesavelino/idea-c2-tool>

⁹<https://spacy.io/api/transformer>

Table 4: Comparison Multi and Single Category of the application of IDEA-C2.

Corpus	Category	BERT Model	Task	Precision	Recall	F1-Score
SciERC	Multicategory	allenai/scibert	NER	65.11%	63.20%	64.14%
			RE	48.27%	20.21%	28.49%
	Singlecategory		NER	76.67%	79.13%	77.88%
			RE	43.68%	26.13%	32.70%
Material Science	Multicategory	roberta-base	NER	79.37%	79.09%	79.23%
			RE	49.76%	29.64%	34.66%
	Singlecategory		NER	70.46%	75.46%	77.41%
			RE	43.28%	40.62%	41.91%

used roberta-base LM because the initial tests' results were superior then when using other LMs.

In the case of SciERC corpus, we can see that the use of the **Singlecategory NER** strategy in training the LM obtained significantly superior results than the **Multicategory NER** strategy, for the three metrics: i) precision: 11.56%; ii) recall: 15.23%; and F1-Score: 13.74%. Even for the RE task, both recall and F1-Score also obtained higher results because the training of the RE task depended on the result of the NER task. On the other hand, in the case of Material Science corpus, the results of the **Multicategory NER** strategy were inferior to the **Singlecategory NER** strategy, but the average loss was 4.8%. In this case, differences in the number of annotated terms or the BERT Model choice may have influenced the results. Varying parameter configurations and other LM model choices may lead to better results, but for the present writing, it was not possible to perform new experiments.

Therefore, based on the results and analysis of this experiment, adopting the **Singlecategory NER** strategy is promising, especially concerning RE results, which showed gains with both corpora. Moreover, there is also the flexibility of the approach that avoids the dependency of pre-defining a set of multiple categories for each domain.

6.2 IDEA-C2-Model Training Evaluation

The second experiment aimed to extend the previous one, and validate the effectiveness of the model training, using the whole set of texts of the Glossary C2 Corpus (BRASIL, 2009). The idea was to evaluate the evolution of the Train Fine-Tuned Language Model sub-process, mentioned in Section 5, and its configuration choices. To this end, the experiment was divided into two stages, initial and final, to analyze and compare the results of the precision, recall, and F1-Score metrics obtained from training the IDEA-C2-Model. The first stage used a previous version of the sub-process, which implemented only $r1$ and $r2$ rules, and that did not fully cover the text of each Glossary

entry. The second stage used the latest version of the sub-process, which implemented the full set of rules and extended the annotation coverage.

To carry out the experiment, the hyperparameters were defined as follows: the Bert model was set to neuralmind/bert-base-portuguese-cased¹⁰ LM (Souza et al., 2020), the Dropout was set to 20% and the vocabulary used was the pt.core_news_sm¹¹ to meet the language of the C2 Glossary corpus (BRASIL, 2009). The remaining hyperparameters were initially assigned their default values. However, for the second stage, there were some adjustments: the spaCy pipeline hyperparameters batch_size were set to 500 and max_length to 100.

Table 5: IDEA-C2-Model training results.

	Ex	Metrics Results		
		Precision	Recall	F1
NER	1	9.93%	17.19%	12.58%
	2	86.56%	86.48%	86.51%
RE	1	0.36%	56.48%	0.72%
	2	98.06%	98.37%	98.21%

Table 5 shows the results of the IDEA-C2-Model training in the two stages, for both NER and RE training tasks. In the first stage, the results obtained in training for the precision, recovery and F1 score metrics were below expectations. In particular, for both tasks, precision was relatively low. However, the recall results, 17.19% for NER task and 56.48% for RE task, confirmed the tendency of a better performance of the **Singlecategory NER** strategy.

In the second stage, the results improved considerably (see Table 5). For the NER task, 22,754 words were processed, with 304 epochs, and LM IDEA-C2-Model obtained about 86% for all metrics. These results confirm that the Singlecategory NER strategy is credited with achieving satisfactory results due to its flexibility and scope. For the RE task, the IDEA-C2-Model training supported by the C2RM sub-specializations was executed in 66 epochs, with a

¹⁰<https://github.com/neuralmind-ai/portuguese-bert>

¹¹<https://spacy.io/models/pt>

threshold value of 0.5 for all evaluation metrics. In this case, it achieved excellent results, reaching 98% for all metrics.

Therefore, this experiment results showed that the pre-annotation and training sub-processes of the IDEA-C2 approach evolved to the point of reaching a very good performance. However, improvements can still be made, such as: improving the pre-annotation task with new rules and replacing the spaCy pipeline to use other existing architectures such as BERT Large. Additionally, new experiments using other C2 corpora may consolidate the initial good performance results.

7 CONCLUSION

This article presented the IDEA-C2, a supervised knowledge graph generation approach supported by a high-level metamodel with Command and Control Relations constructs, called C2RM. This metamodel provides high flexibility to the approach since the domain entities categories are not prefixed. In the experiments carried out, promising results were obtained, achieving more than 70% precision and recall in the training of the LM based on the corpus from other published works. The approach uses distance supervision methods to pre-annotate Command and Control Doctrinal Text for model fine-tuning. Likewise, the implemented IDEA-C2-Model application showed remarkable results in training NER and RE models, achieving over 80% precision and 98% recall, using as input the Glossary C2 corpus. Finally, these experiments using the IDEA-C2-Tool proved the usefulness and feasibility of the proposed approach and it is already able to generate the IDEA-C2-KG, which is available for queries and inferences. Future work includes improving pre-annotation tasks and evaluating entity and relation categories statistically.

ACKNOWLEDGEMENTS

This research has been funded by FINEP/DCT/FAPEB (no. 2904/20-01.20.0272.00) under the S2C2 project.

REFERENCES

- Augenstein, I., Das, M., Riedel, S., Vikraman, L., et al. (2017). ScienceIE - Extracting keyphrases and relations from Scientific Publications. In *Proc Int Work on Semantic Evaluation*, pages 546–555, Canada. ACL.
- BRASIL (2009). Glossário de Termos e Expressões para uso no Exército. *Estado Maior do Exército*.
- Chaudhri, V. K., Cheng, B., Overholtzer, A., et al. (2013). Inquire biology: A textbook that answers questions. *AI Magazine*, 34(3):55–72.
- Dang, L. D., Phan, U. T., and Nguyen, N. T. (2023). GENA: A knowledge graph for nutrition and mental health. *Journal of Biomedical Informatics*, 145:104460.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc the Conf of the North American Chapter of the ACL: Human Language Technologies, Volume 1*, pages 4171–4186, Minnesota. ACL.
- Hogan, A., Blomqvist, E., Cochez, M., et al. (2021). Knowledge Graphs. *ACM Computing Surveys*, 54(4).
- Kent, W. (2012). *Data and Reality: A Timeless Perspective on Perceiving and Managing Information*. Technics publications.
- Lee, J., Yoon, W., Kim, S., Kim, D., et al. (2019). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.
- Liu, P., Qian, L., Zhao, X., and Tao, B. (2023). The construction of knowledge graphs in the aviation assembly domain based on a joint knowledge extraction model. *IEEE Access*, 11:26483–26495.
- Luan, Y., He, L., Ostendorf, M., and Hajishirzi, H. (2018). Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proc Conf on Empirical Methods in NLP*, pages 3219–3232, Brussels, Belgium. ACL.
- Mintz, M., Bills, S., Snow, R., and Jurafsky, D. (2009). Distant supervision for relation extraction without labeled data. In *Proc of the Joint Conf of the 47th Annual Meeting of the ACL and the Int Joint Conf on NLP of the AFNLP*, pages 1003–1011, Singapore. ACL.
- Russell, S. and Norvig, P. (2010). *Artificial Intelligence: A Modern Approach*. 3ed. Prentice Hall.
- Souza, F., Nogueira, R., and Lotufo, R. (2020). BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In Cerri, R. and Prati, R. C., editors, *Intelligent Systems*, pages 403–417, Cham. Springer Int Pub.
- Spala, S., Miller, N., Derroncourt, F., and Dockhorn, C. (2020). SemEval-2020 task 6: Definition extraction from free text with the DEFT corpus. In *Proc of the Fourteenth Workshop on Semantic Evaluation*, pages 336–345, Barcelona. ICCL.
- Weston, L., Tshitoyan, V., Dagdelen, J., Kononova, O., et al. (2019). Named entity recognition and normalization applied to large-scale information extraction from the materials science literature. *Journal of Chemical Information and Modeling*, 59(9):3692–3702.
- Zhao, Q., Huang, H., and Ding, H. (2021). Study on military regulations knowledge construction based on knowledge graph. In *2021 7th Int Conf on Big Data and Information Analytics (BigDIA)*, pages 180–184.
- Zhou, J., Li, X., Wang, S., and Song, X. (2022). NER-based military simulation scenario development process. *Journal of Defense Modeling and Simulation*.