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

Jones O. Avelino[1,2] [a], Giselle F. Rosa[1] [b], Gustavo R. Danon[1] [c], Kelli F. Cordeiro[3] [d]
and Maria Cláudia Cavalcanti[1] [e]

[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