IDEA-C2: uma abordagem híbrida de modelagem conceitual apoiada por um modelo de linguagem e um metamodelo de dados no contexto de Comando e Controle

Jones O. Avelino[1,2] **Kelli F. Cordeiro**[3], **Maria Cláudia Cavalcanti**[1]

[1]Pós-graduação em Engenharia de Defesa – IME – Rio de Janeiro – RJ – Brasil

[2]Centro de Análise de Sistemas Navais – Rio de Janeiro – RJ – Brasil

[3]Subchefia de Comando e Controle – Ministério da Defesa – Brasília – DF – Brasil

jones.avelino@marinha.mil.br, kelli.cordeiro@defesa.gov.br, yoko@ime.eb.br

Abstract. *This thesis proposes IDEA-C2, a hybrid supervised approach that combines doctrinal texts, semantic resources, and a high-level metamodel to annotate corpora and fine-tune language models (LM). Unlike data-driven approaches, which build subsymbolic domain models (DMs), and theory-driven approaches, which face challenges in extracting classes and relationships from texts, IDEA-C2 employs heuristic pre-annotation and generates flexible knowledge graphs (KGs) for exploratory queries and inference, supporting the construction of the DM. Evaluated in six experiments, the IDEA-C2 approach achieved 95% precision in classes and 76% in relations during pre-annotation, as well as precision and recall above 85% in context-adapted LM. In an experiment with 28 participants, 40% of the KG's classes and relationships were found to be similar to DMs constructed using traditional methods. The results demonstrate the usefulness and feasibility of IDEA-C2 in generating artifacts and in practical applications.*

Resumo. *A tese propõe o IDEA-C2, uma abordagem supervisionada híbrida que combina textos doutrinários, Recursos Semânticos e um metamodelo de alto nível para anotar corpora e ajustar Modelos de Linguagem (ML). Diferente das abordagens orientadas por dados (data-driven), que constroem Modelos de domínio (DM) subsimbólicos, e das orientadas por teoria (theory-driven), que enfrentam desafios na extração de classes e relações de textos, o IDEA-C2 emprega pré-anotação heurística e gera Knowledge Graphs (KG) flexíveis para consultas exploratórias e realização de inferências, apoiando a construção do DM. Avaliada em seis experimentos, a abordagem IDEA-C2 alcançou 95% de precisão em classes e 76% em relações na pré-anotação, além de uma precisão e cobertura acima de 85% no ML ajustado ao contexto. Em experimento com 28 participantes, 40% das classes e relações do KG se mostraram similares aos DM construídos de modo tradicional. Os resultados demonstram a utilidade e viabilidade do IDEA-C2 na geração de artefatos e na aplicação prática.*

1) Dados e Destaques da Defesa de Tese: Categoria: D.Sc.; **Defendida em:** 09/02/2026

Orientador: Maria Cláudia Cavalcanti (IME); Coorientador: Kelli F. Cordeiro (IME);

Membros da banca: Ronaldo Ribeiro Goldschmidt (IME); Julio Cesar Duarte (IME); João Luiz Rebelo Moreira (UT - Holanda); Sérgio Manuel Serra da Cruz (UFRJ).

Destaques:

- Propusemos uma abordagem híbrida para apoiar a construção de modelos de domínio, combinando as abordagens *theory-driven* e *data-driven*, comparativamente à construção que utiliza essencialmente a abordagem *theory-driven*.
- Realizamos um experimento comparativo de construção de modelos de domínio, utilizando nossa abordagem, envolvendo 28 pessoas, reunindo perfis e experiências distintas;
- Analisamos os resultados da análise comparativa de construção de modelos de domínio (DM), demonstrando que 40% das classes e relações do Knowledge Graph (KG) foram similares às dos DM construídos de maneira tradicional.
- Propusemos uma abordagem de pré-anotação de corpora, baseado em um metamodelo de dados conceitual combinado com um recurso semântico, que gerou um *dataset* contendo mais de 3 mil textos anotados no domínio militar.
- O *dataset* resultante do pré-anotador alcançou uma precisão de 95% nas entidades e 76% nas relações, culminando em um modelo de linguagem ajustado ao contexto militar com uma precisão e cobertura acima de 85%.
- Submetemos a nossa abordagem de pré-anotação a corpus distintos e

Tese disponível em: `http://bit.ly/4hliqtg`

Publicações disponíveis em: `https://bit.ly/4fOrAfN`

Código-fonte e dados disponíveis em: `https://bit.ly/3TbyILo`

2) Contexto e problema: Os avanços recentes em Processamento de Linguagem Natural (PLN) têm se destacado nas tarefas de Reconhecimento de Entidades Nomeadas (NER) e Extração de Relações (RE). A obtenção de conhecimento a partir de dados textuais foi impulsionada pelo avanço dos Modelos de Linguagem (ML), cujo desempenho pode ser aprimorado por meio do ajuste fino, ou *fine-tuning*, em domínios específicos [Luan et al. 2018]. Contudo, abordagens orientadas por dados, ou *Data-Driven* (DD), geram modelos subsimbólicos que apresentam limitações, como falta de explicabilidade e vieses em função de sua aleatoriedade estocástica [Saba 2023]. Em contraste, abordagens orientadas por teoria, ou *Theory-Driven* (TD), baseiam-se em conceituações formais para a construção de Modelos de Domínio (DM) simbólicos, embora enfrentem desafios na extração de classes e relações relevantes a partir de textos [Guizzardi et al. 2023].

Embora ML ajustados em um determinado domínio, como SciERC [Luan et al. 2018], BioBERT [Lee et al. 2019] e TForMIX [Rosa et al. 2025], possam aprimorar o fine-tuning, eles ainda são escassos e geralmente limitados a categorias previamente definidas. O ajuste fino de ML também enfrenta desafios como a necessidade de corpora extensos, representativos e anotados, além de curados e validados. Entretanto, esses conjuntos de dados também são escassos, especialmente em domínios específicos, como, por exemplo, o militar. Para contornar esse problema, alguns estudos utilizam artigos científicos (e.g. SciERC que é formado por textos da área de científica [Luan et al. 2018]), além de livros didáticos e documentos doutrinários na construção de corpora devido ao seu caráter informativo e pedagógico [Liu et al. 2023, Zhao et al. 2021]. Além disso, há autores que usam ML pré-treinados no idioma em que são aplicados para obter melhores resultados (e.g. BERTimbau [Souza et al. 2020] que utiliza textos em língua portuguesa). Outro desafio é o custo de anotação de corpora, especialmente em textos brutos. Para

reduzir esse custo, alguns estudos utilizam métodos supervisionados à distância [Mintz et al. 2009] com apoio de recursos semânticos, como ontologias, vocabulários e glossários [Zhou et al. 2022, Fries et al. 2021]. Apesar dos resultados promissores, essas abordagens ainda são limitadas a domínios específicos e a estruturas ontológicas previamente definidas.

Dessa forma, os métodos atuais são limitados a um conjunto de categorias de entidades e de relações específicas, dificultando a sua aplicação em diferentes domínios. Adicionalmente, há autores que propõem a utilização de Knowledge Graph (KG) para estruturar os dados utilizados no ajuste do ML, permitindo que os usuários realizem consultas, inferências e até enriquecimento de *datasets* [Hogan et al. 2021, Yang et al. 2022]. Porém, os métodos de construção de KG se baseiam em categorias restritas, limitando a sua aplicação em domínios específicos e dificultando o alcance de resultados mais abrangentes.

3) Objetivo: Este tese tem como objetivo especificar uma abordagem híbrida, combinando as abordagens DD e TD, para apoiar a construção de um DM. Essa abordagem é apoiada por um KG gerado através de um ML ajustado, em especial, nas tarefas de NER e RE. O corpus utilizado no ajuste do ML tem como base os dados textuais não-estruturados de documentos doutrinários do contexto militar. Nesse sentido, este trabalho explora algumas lacunas, tais como: a) abordagens de ajuste fino de ML com categorias restritas para aplicações em multidomínio; b) geração de KG como estruturação e exploração de dados a partir de textos; e c) apoio à construção de DM a partir de interações com o ML.

4) Resumo da solução proposta: Este trabalho propõe o IDEA-C2 (generat**I**on of knowle**D**ge graphs bas**E**d on **A**rtificial intelligence of **C2** Domain), uma abordagem supervisionada híbrida que combina textos doutrinários, recursos semânticos e um metamodelo de alto nível (C2RM) para anotar corpora e ajustar modelos de linguagem pré-treinados em língua portuguesa. A abordagem emprega técnicas de pré-anotação heurística e permite a geração de knowledge graphs (KG) flexíveis, viabilizando consultas exploratórias e inferências, a fim de apoiar o desenvolvimento de modelos de domínio (DM), combinando as abordagens DD e TD.

5) Avaliação dos resultados: A abordagem IDEA-C2 foi avaliada em seis experimentos distintos, obtendo resultados promissores. Na pré-anotação do corpus, IDEA-C2 alcançou uma precisão de 95% nas entidades e 76% nas relações, culminando em um ML ajustado ao contexto com uma precisão e cobertura acima de 85%. Em outro experimento mais amplo, envolvendo 28 participantes, ao aplicar o ML ajustado combinado com o KG no apoio à construção de um DM, os resultados do IDEA-C2 mostraram que 40% das classes e relações do KG foram similares às dos DM construídos de maneira tradicional. Esses resultados demonstram a utilidade e viabilidade da abordagem IDEA-C2 tanto na geração de artefatos essenciais ao ajuste de um ML e geração de KG quanto na sua aplicação na construção de um DM.

6) Contribuições e avanços em relação ao estado da arte: Nesta tese, foi realizado um estudo que comparou a construção tradicional de DM com o apoio de nossa abordagem híbrida, que combinou as abordagens DD (subsimbólica) e TD (simbólica), através das interações com o ML ajustado e o KG. O estudo identificou que o uso isolado de ambas as abordagens possuem limitações, influenciando a construção restrita de um DM. Nesse

sentido, os resultados do estudo demonstraram que a adoção de uma abordagem híbrida poderia se valer tanto dos padrões estatísticos do ML quanto da conceituação formal do domínio. Porém, um dos desafios foi encontrar corpora, KG e ML ajustados ao domínio militar, em especial na área de Comando e Controle (C2). Dada a escassez de corpora anotados e o custo de anotação, foi elaborado um processo de pré-anotação, combinando as metacategorias de um metamodelo conceitual de dados com regras heurísticas, explorando expressões regulares. Com base nesse processo de pré-anotação foi construído o protótipo PREAnoTeTool para realização de experimentos. Com isso, buscou-se minimizar a anotação e curadoria, bem como apoiar a geração de ML ajustados no domínio alvo. Paralelamente, aproveitou-se dos dados utilizados no ajuste do ML, das metacategorias do metamodelo e da extração do conhecimento do ML para definir um processo semiautomatizado de geração de KG. Aliado a isso, foi construído o protótipo IDEA-C2-Tool para realizar os experimentos e interações com o ML e o KG. De modo geral, os resultados dos experimentos a que o IDEA-C2 foi submetido, demonstraram que a nossa abordagem supera os métodos tradicionais, evidenciando que uma abordagem híbrida pode agregar maior flexibilidade, reduzir custos operacionais e favorecer a construção de um DM no domínio militar, ou até apoiar a revisão de modelagens já existentes com o objetivo de aumentar a interoperabilidade entre os aplicativos de C2.

7) Produção Científica: Nosso trabalho foi publicado em simpósios e submetido a journal da área de tratamento de dados e PLN:

I. Simpósios

1. Avelino, Rosa, et al. "**Knowledge graph generation from text using supervised approach supported by a relation metamodel: An application in c2 domain**" na 26ª International Conference on Enterprise Information Systems (ICEIS).
2. Avelino, Rosa, et al. "**PREAnoTe: Uma abordagem de anotação de corpus para o ajuste fino de Large Language Model pré-treinado**" no XXXIX Simpósio Brasileiro de Bancos de Dados (SBBD).

II. Em avaliação

3. Avelino, Cordeiro, Cavalcanti "**Towards a hybrid conceptual modeling approach combining a fine-tuned Language Model and a metamodel**" ao Journal IEEE Computer Society.

III. Produção Técnica, registros e patentes

4. O prototótipo PREAnoTeTool obteve seu registro de programa de computador no Instituto Nacional de Propriedade Industrial (INPI), em 15ABR2025, através do processo BR 51 2025 001402-3.

8) Prêmios e Conquistas: Apesar de não haver uma premiação. No ICEIS de 2024, o artigo "**Knowledge graph generation from text using supervised approach supported by a relation metamodel: An application in C2 domain**" foi um dos destaques da conferência, sendo convidado para uma versão ampliada a ser publicada no livro Lecture Notes in Business Information Processing (LNBIP).

Agradecimentos

Agradecemos ao CNPq - Conselho Nacional de Desenvolvimento Científico e Tecnológico e à FINEP/DCT/FAPEB (nº 2904/20-01.20.0272.00), pelo apoio ao Projeto 'Sistemas de Sistemas de Comando e Controle".

Referências

Fries, J. A., Steinberg, E., Khattar, S., Fleming, S. L., Posada, J., Callahan, A., and Shah, N. H. (2021). Ontology-driven weak supervision for clinical entity classification in electronic health records. *Nature communications*, 12(1):2017.

Guizzardi, G., Pastor, O., and Storey, V. C. (2023). Thinking Fast and Slow in Software Engineering . *IEEE Software*, 40(06):139–142.

Hogan, A., Blomqvist, E., Cochez, M., et al. (2021). Knowledge graphs. *ACM Computing Surveys*, 54(4).

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (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 *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232, Brussels, Belgium. Association for Computational Linguistics.

Mintz, M., Bills, S., Snow, R., and Jurafsky, D. (2009). Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, Suntec, Singapore. Association for Computational Linguistics.

Rosa, G. F., Avelino, J. O., Cavalcanti, M. C., and Duarte, J. C. (2025). TForMIX: A method that combines LLM and Multidimensional Modeling for Technological Foresight. *IEEE Access*, 13:153320–153339.

Saba, W. S. (2023). Stochastic llms do not understand language: Towards symbolic, explainable and ontologically based llms. In Almeida, J. P. A., Borbinha, J., Guizzardi, G., Link, S., and Zdravkovic, J., editors, *Conceptual Modeling*, pages 3–19, Cham. Springer Nature Switzerland.

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 International Publishing.

Yang, J., Han, S. C., and Poon, J. (2022). A survey on extraction of causal relations from natural language text. *Knowledge and Information Systems*, 64(5):1161–1186.

Zhao, Q., Huang, H., and Ding, H. (2021). Study on military regulations knowledge construction based on knowledge graph. In *2021 7th International Conference 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. *The Journal of Defense Modeling and Simulation*, 20(4):563–575.