Desafios dos Grafos de Conhecimento: uma proposta de avaliação de sistemas OpenIE.

Samuel Rios da Silva[1], Aline Athaydes[1], Babacar Mane[1], Daniela Barreiro Claro[1], Marlo Souza[1], Fernando H. de A. Moraes Neto [1], Larrissa Dantas[1], Rerisson Cavalcante[1]

[1]FORMAS Research Center on Data and Natural Language
Institute of Computing – Federal University of Bahia (UFBA)
Av. Milton Santos, s/n - Campus de Ondina – 40.170-110 – Salvador – BA – Brazil

{samuelrs,alineathaydes,babacarm,dclaro,msouza1}@ufba.br

{fernando.humberto,larrissasilva}@ufba.br

Abstract. *Open Information Extraction (OpenIE) faces challenges in evaluating its models. With the use of traditional metrics from other areas and the absence of a gold standard corpus, the difficulty arises in evaluating all possible extractions generated by the models. In this work, we propose a comparative evaluation method for different OpenIE models focused on the Portuguese language using knowledge graphs. The results obtained show that models capable of generating a greater number of accurate triples tend to deliver better performance, highlighting the importance of balancing quantity and quality in the OpenIE task.*

Resumo. *A Extração de Informação Aberta (OpenIE) enfrenta desafios na avaliação de seus modelos. Com a utilização de métricas tradicionais de outras áreas e a ausência de um corpus de ouro, surge a dificuldade de avaliar todas as possíveis extrações geradas pelos modelos. Neste trabalho, propomos um método de avaliação comparativa entre diferentes modelos de OpenIE voltados para a língua portuguesa utilizando grafos de conhecimento. Os resultados obtidos demonstram que modelos capazes de gerar um maior número de triplas com precisão tendem a oferecer melhor desempenho, evidenciando a importância de equilibrar quantidade e qualidade na tarefa de OpenIE.*

1. Introdução

A Extração de Informação Aberta (do Inglês, *Open Information Extraction*), estuda métodos computacionais para identificar informações semânticas estruturadas de fontes não estruturadas [Glauber and Barreiro Claro 2018]. As estruturas semânticas são representadas por tuplas relacionais consistindo de um conjunto de argumentos e uma expressão denotando uma relação semântica entre eles no formato ⟨ arg1; rel; arg2 ⟩ [Claro et al. 2024].

A avaliação desses sistemas, em geral, tem se apoiado em métricas tradicionais da área de Recuperação de Informação, como Precisão, Recall e sua média harmônica, a pontuação F1, aplicadas sobre um corpus ouro. Esse corpus é formado por sentenças cujas extrações foram validadas por especialistas, servindo como referência para medir o desempenho dos modelos. No entanto, essas métricas quantitativas isoladas não são suficientes para capturar a qualidade semântica das tuplas extraídas, principalmente diante dos desafios específicos da tarefa. Entre esses desafios, destacam-se:

- Não há um guia de anotação de corpus padrão para extrair tuplas em uma tarefa de Extração de Informação Aberta.
- Não existe um corpus suficientemente grande em português para Extração de Informação Aberta.
- Os métodos de Extração de Informação Aberta apresentam diferentes características de extração. Por exemplo, alguns extraem apenas a relação mínima, enquanto outros extraem a relação composta. Essas diferenças podem resultar em tuplas distintas mesmo quando aplicadas à mesma sentença.
- Divergência entre os métodos de Extração de Informação produzem incoerências nos resultados devidos às métricas quantitativas binárias não avaliarem os subconjuntos de unidades lexicais existentes entre Argumentos e Relações, tornando as métricas quantitativas divergentes para comparação das triplas.

Neste contexto, este trabalho propõe uma abordagem de avaliação dos métodos de Extração de Informação Aberta com foco na análise de tuplas semânticas extraídas por abordagens distintas, como modelos simbólicos e neurais. Para isso, cada conjunto de tuplas gerado a partir das sentenças é convertido em um grafo de conhecimento, estrutura que representa entidades e suas respectivas relações por meio de nós e arestas [Paulheim 2017]. A avaliação baseia-se na análise quantitativa desses elementos, bem como na similaridade estrutural e na sobreposição semântica entre os grafos desenvolvidos.

2. A Metodologia

Uma vez que as extrações são feitas, geramos o grafo de conhecimento respectivo para o conjunto de extrações do modelo OpenIE. Por exemplo, a partir da sentença "*Salvador é uma cidade bela e possui lindas praias.*", podemos ter os seguintes conjuntos de extrações:

Conjunto 1:

i) **arg1**: Salvador; **relação**: é; **arg2**: uma cidade.
ii) **arg1**: Salvador; **relação**: é; **arg2**: uma cidade bela.
iii) **arg1**: Salvador; **relação**: possui; **arg2**: lindas praias.

Conjunto 2:

i) **arg1**: Salvador; **relação**: é; **arg2**: uma cidade.

A Figura 1 mostra os grafos de conhecimento gerados a partir de cada conjunto de extrações. Assim, podemos comparar tais grafos em termos da quantidade de vértices, grau médio, isomorfimo (considerando os labels dos nós e arestas), subgrafos e grafos vazios.

Mais precisamente, dada uma única extração na forma $\langle$ arg1; rel; arg2 $\rangle$, os argumentos 1 e 2 (arg1 e arg2) da tripla se tornarão vértices no grafo de conhecimento e a relação (rel) se tornará uma aresta conectando tais argumentos. Um grafo com maior número de vértices indica que o modelo possui uma capacidade maior de reconhecimento de entidades distintas.

A partir da Figura 1, no grafo gerado a partir do conjunto 1, é possível perceber que a entidade *"Salvador"* possui um maior grau do que no grafo gerado a partir do conjunto 2. Isso se dá ao fato de que existem mais extrações das quais tal entidade participa. Assim, uma vez que o grau médio é dado pela razão entre a soma dos graus dos vértices em G e a quantidade de vértices em G, um grafo com maior grau médio indica que o modelo de extração possui uma maior capacidade de extrair múltiplos fatos sobre uma mesma entidade.

Podemos também saber quantas entidades distintas foram extraídas de uma dada sentença por meio da quantidade de vértices do grafo, uma vez que, mesmo aparecendo em múltiplas extrações, como é o caso de *"Salvador"* no conjunto 1, tal entidade será representada por um único vértice.

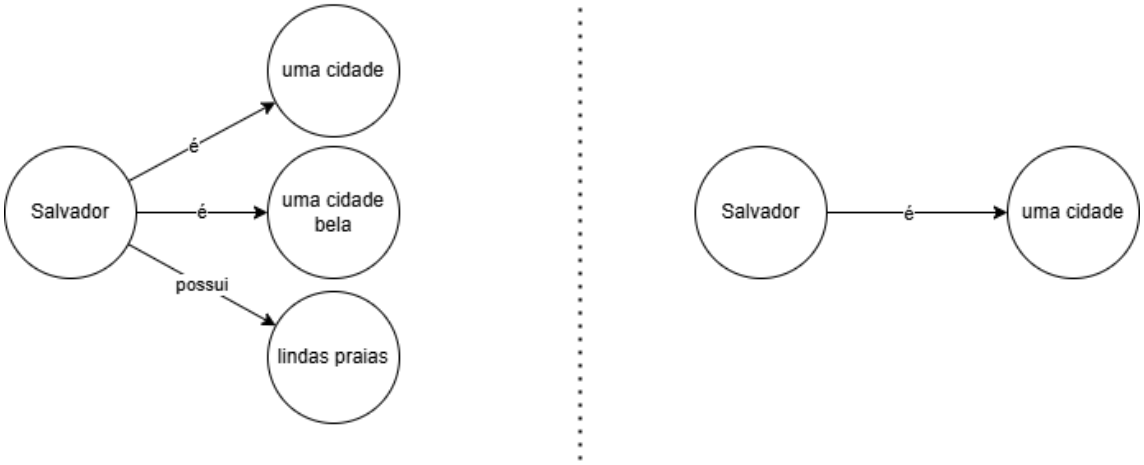

Figure 1. Grafos gerados pelas extrações do Conjunto 1 (lado esquerdo) e do Conjunto 2 (lado direito).

Partindo do fato de que conjuntos de extrações diferentes produzem grafos de conhecimento distintos, comparando a estrutura dos grafos de conhecimento e sendo eles isomorfos (considerando a igualdade dos labels dos seus nós e arestas), significa que os modelos produziram o mesmo conjunto de extrações para a mesma sentença. No entanto, nos casos em que eles não são isomorfos, podemos avaliar se um grafo é subgrafo do outro, o que pode nos indicar que o modelo do grafo maior obteve uma maior quantidade de fatos em suas extrações.

Podemos também fazer a comparação das classes gramaticais de cada palavra presente em cada elemento da tripla também por meio de grafos a fim de identificar padrões. Para isso, para cada grafo gerado a partir de um conjunto de extrações, podemos criar um novo grafo com cada nó e aresta nomeados com as classes gramaticais do label no grafo original. Assim, podemos aplicar as mesmas comparações em termos da quantidade de vértices, grau médio, isomorfimo, subgrafos e grafos vazios.

3. Experimento e Desafios

Para a construção e comparação dos grafos, seguimos algumas etapas essenciais. O primeiro passo foi a adaptação do corpus BIA [Queiroz et al. 2023], que possui um conjunto de 360 sentenças e suas extrações validadas por humanos. Nessa etapa, extraímos as sentenças do corpus e criamos um arquivo csv com cada linha contendo uma sentença e seu identificador. Com isso, submetemos as sentenças para cada um dos modelos - neural e baseado em regra - para que as extrações fossem feitas para cada sentença.

A próxima etapa foi fazer o mapeamento de cada palavra presente na extração com a sua classe gramatical, dada a sentença original. Daí, então, surge o primeiro desafio. Foi observado que as expansões de algumas preposições, por exemplo "do" para "$de\ o$", se diferenciavam entre os modelos. Tivemos então que mapear tais casos e tratar de forma que tais diferenças não impactassem nos resultados.

Um segundo desafio, ainda nessa etapa, foi a necessidade de fazer a desambiguação das palavras que se repetiam na sentença, mas que não tinham a mesma classe gramatical. Por exemplo, na sentença "*Os títulos lastreados em hipotecas comerciais, os lastreados em ativos e os CDOs atingiram o pico em 2006.*", a palavra "lastreados" aparece duas vezes na sentença e tais ocorrências possuem classes gramaticais diferentes.

Como solução, definimos uma janela de contexto de tamanho que varia de 1 a 7 tokens à frente, de modo que possibilite a identificação da ocorrência da palavra repetida que aparece no argumento ou na relação da tripla. Para os casos em que, mesmo assim, não foi possível fazer a desambiguação, selecionamos, aleatoriamente, uma das classes gramaticais. Com as extrações e suas classes gramaticais atribuídas, então, foram construídos os grafos respectivos.

4. Resultados parciais e Discussões

Das 360 sentenças presentes no corpus BIA, ignoramos 73 sentenças, pois algumas delas tinham caracteres que os modelos não conseguiam processar ou porque tivemos problemas no momento da definição das classes gramaticais. Assim, para cada uma das 287 sentença restantes, construímos 3 grafos de conhecimento: cada um a partir das extrações do DPTOIE [Oliveira et al. 2023], do PortNOIE [Cabral et al. 2022] e do corpus BIA.

Em relação aos grafos gerados a partir do BIA, tanto o PortNOIE quanto o DPTOIE tiveram apenas 3 grafos isomorfos. No entanto, em relação aos subgrafos, enquanto apenas 2 grafos do DPTOIE são subgrafos do BIA, 85 grafos do PortNOIE são subgrafos do BIA. Por outro lado, 81 grafos do BIA são subgrafos do DPTOIE e 79 do BIA são subgrafos do PortNOIE.

Das 287 sentenças processadas, o PortNOIE não gerou extrações para aproximadamente 20,5% (59) delas, e o DPTOIE não gerou extrações para aproximadamente 3,4% (10) das sentenças, o que resultou em grafos vazios. Em relação ao grau médio dos grafos, percebemos que o DPTOIE possui os maiores graus médios, variando de 1 a 2,5, em contraste com o PortNOIE que possui os valores mais baixos para essa métrica, com apenas 8 grafos com grau médio maior que 1. Entre eles, o BIA possui um grau médio variando de 1 a 1,8, mas cuja maioria tinha grau médio 1.

5. Conclusão

Os resultados mostram o desempenho superior do DPTOIE no que se refere ao reconhecimento de múltiplos fatos associados a uma mesma entidade, dado que o grau médio de seus grafos é maior em comparação ao PortNOIE. A grande quantidade de grafos vazios gerados pelo PortNOIE mostra sua limitação no que se refere à cobertura das sentenças.

Podemos observar que os modelos possuem uma grande variação na extração de suas triplas, distinguindo-se entre eles, dado que poucos grafos eram isomorfos entre si e menos da metade dos pares de grafos analisados eram subgrafos um do outro.

A abordagem proposta reforça a importância de métodos de avaliação que considerem não apenas a quantidade de triplas extraídas, mas também a qualidade e a riqueza estrutural das informações representadas.

Agradecimentos

Ao Escavador e a FAPESB por meio do projetos TIC 0002/2015, CCE 0022/2023 e INCITE PIE0002/2022.

References

Cabral, B., Souza, M., and Claro, D. B. (2022). Portnoie: A neural framework for open information extraction for the portuguese language. In *International Conference on Computational Processing of the Portuguese Language*, pages 243–255. Springer.

Claro, D. B., Santos, J., Souza, M., Vieira, R., and Pinheiro, V. (2024). Extração de informação. In Caseli, H. M. and Nunes, M. G. V., editors, *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, book chapter 20. BPLN, 2 edition.

Glauber, R. and Barreiro Claro, D. (2018). A systematic mapping study on open information extraction. *Expert Systems with Applications*, 112:372–387.

Oliveira, L., Claro, D. B., and Souza, M. (2023). Dptoie: a portuguese open information extraction based on dependency analysis. *Artificial Intelligence Review*, 56(7):7015–7046.

Paulheim, H. (2017). Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic web*, 8(3):489–508.

Queiroz, B., Cavalcante, R., and Claro, D. (2023). Desafios da tarefa de extração de informação aberta: uma abordagem metodológica de um corpus automatizado até o corpus manual. In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 388–392, Porto Alegre, RS, Brasil. SBC.