UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO DE COMUNICAÇÃO E EXPRESSÃO
DEPARTAMENTO DE LÍNGUA E LITERATURA VERNÁCULAS
CURSO DE LETRAS – LÍNGUA PORTUGUESA E LITERATURAS

Manuela Pacheco de Andrade

Uso da engenharia de *prompt* em LLMs para a resolução de correferência na língua portuguesa

Florianópolis
2024

Manuela Pacheco de Andrade

Uso da engenharia de *prompt* em LLMs para a resolução de correferência na língua portuguesa

Trabalho de Conclusão de Curso submetido ao curso de Letras – Língua Portuguesa e Literaturas do Centro de Comunicação e Expressão da Universidade Federal de Santa Catarina como requisito parcial para a obtenção do título de Bacharela em Letras – Língua Portuguesa e Literaturas.

Orientador: Prof. Dr. Alckmar Luiz dos Santos
Coorientador: Prof. Dr. Renato Fileto

Florianópolis
2024

Andrade, Manuela Pacheco de
Uso da engenharia de prompt em LLMs para a resolução de correferência na língua portuguesa / Manuela Pacheco de Andrade ; orientador, Alckmar Luiz dos Santos, coorientador, Renato Fileto, 2024.
96 p.

Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Santa Catarina, Centro de Comunicação e Expressão, Graduação em Letras - Língua Portuguesa, Florianópolis, 2024.

Inclui referências.

1. Letras - Língua Portuguesa. 2. Resolução de Correferência. 3. Engenharia de Prompt. 4. Processamento de Língua Natural. I. Santos, Alckmar Luiz dos. II. Fileto, Renato. III. Universidade Federal de Santa Catarina. Graduação em Letras - Língua Portuguesa. IV. Título.

Manuela Pacheco de Andrade

Uso da engenharia de *prompt* em LLMs para a resolução de correferência na língua portuguesa

Este Trabalho de Conclusão de Curso foi julgado adequado para obtenção do título de Bacharela em Letras – Língua Portuguesa e Literaturas e aprovado em sua forma final pelo Curso Letras – Língua Portuguesa e Literaturas.

Florianópolis, 01 de novembro de 2024.

Documento assinado digitalmente
ALEXANDRE ANDRE NODARI
Data: 18/11/2024 18:45:33-0300
CPF: ***.893.189-**
Verifique as assinaturas em https://v.ufsc.br

Coordenação do curso

Banca examinadora

Documento assinado digitalmente
Alckmar Luiz dos Santos
Data: 18/11/2024 14:33:40-0300
CPF: ***.782.368-**
Verifique as assinaturas em https://v.ufsc.br

Prof. Dr. Alckmar Luiz dos Santos
Orientador(a)

Documento assinado digitalmente
ROBERLEI ALVES BERTUCCI
Data: 18/11/2024 16:18:23-0300
CPF: ***.474.429-**
Verifique as assinaturas em https://v.ufsc.br

Prof. Dr. Roberlei Alves Bertucci
Universidade Tecnológica Federal do Paraná

Documento assinado digitalmente
Carina Friedrich Dorneles
Data: 18/11/2024 17:26:21-0300
CPF: ***.834.250-**
Verifique as assinaturas em https://v.ufsc.br

Profa. Dra. Carina Friedrich Dorneles
Universidade Federal de Santa Catarina

Florianópolis, 2024.

Ao Davi, que estaria muito feliz com isso.

AGRADECIMENTOS

Agradeço ao meu pai, que ficou quase tão interessado no tema desse trabalho quanto eu. À minha mãe, que torna tudo infinitamente mais leve, e à minha família toda, que é tão especial.
Um agradecimento especial à dupla Jorge e Eli, por me guiarem no início dessa pesquisa, e ao Professor Popel, cuja disposição em me ajudar foi vital para a progressão do meu trabalho.
Também agradeço ao Professor Marco Antonio, que me apresentou ao tema, e ao meu orientador e coorientador, Professor Alckmar e Professor Renato Fileto, por toda a ajuda.
Um abraço gigante para o Nicolas, e outro para o Aluá. Não sei nem como dizer o quanto vocês colorem a minha vida e o quanto vocês me ajudaram nesse processo todo.
E outros vários abraços para todos os meus amigos que fazem a vida valer a pena: Vargas, Maria, Artur, Orti e Letícia.
E para a Dudu, que não soube se ia colocar na parte da família ou dos melhores amigos, então coloquei aqui embaixo.

RESUMO

Em condições normais de comunicação, elementos linguísticos se articulam em um conjunto coeso, integrado a informações extralinguísticas, formando o que se denomina discurso. A correferência, junto da anáfora – com a qual frequentemente coocorre –, é um fenômeno indissociável da construção dessa estrutura, permeando toda comunicação linguística. Dessa forma, no contexto do Processamento de Língua Natural, a resolução de correferência (RC) é essencial para a compreensão do discurso, contribuindo para o acesso à complexa rede de conexões subjacente a ele. Trata-se de uma tarefa desafiadora, que depende de habilidades linguísticas e extralinguísticas – como compreensão do contexto textual e conhecimento de mundo – e que facilita a execução de diversas outras tarefas de Processamento de Língua Natural. Atualmente, grande parte dos modelos de resolução de correferência com os melhores resultados baseia-se no aprendizado de máquina supervisionado. No entanto, o desempenho dessa abordagem depende altamente de corpora anotados de grande volume, recurso oneroso e escasso no português. Diante desse cenário, o uso de engenharia de *prompt* em grandes modelos de linguagem pré-treinados se mostra como potencial alternativa menos onerosa aos modelos com o uso de aprendizado supervisionado, visto que dispensa a necessidade de dados anotados para treinamento. Este estudo avalia o desempenho do uso dessa estratégia para a RC de entidade no português por dois modelos de linguagem – o ChatGPT-4o e uma versão do Chat-GPT4 customizada por meio do uso de um *system prompt* e de arquivos de conhecimento. Para os testes, foram elaboradas duas séries de *prompts zero* e *few-shot*. Além disso, foi realizada a revisão de um segmento do corpus Summ-it++ – segundo maior corpus com informação de correferência do português – com base nas diretrizes do corpus OntoNotes, gerando uma versão revisada que foi utilizada, juntamente com a versão original, como corpora de teste. Ambos os corpora foram harmonizados conforme o formato CorefUD, possibilitando a avaliação dos resultados com o uso da ferramenta de avaliação CorefUD scorer, que engloba as principais métricas da área. No CoNLL Score os testes alcançaram um desempenho 17,75% inferior ao do modelo estado da arte na resolução do corpus Summ-it++. Embora não tenha atingido níveis de estado da arte, a abordagem se mostrou promissora e demonstrou um desempenho que, ainda que inferior, é relevante em relação ao dos melhores modelos na resolução do mesmo corpus, sugerindo potencial para futuros aprimoramentos.

Palavras-chave: resolução de correferência; engenharia de *prompt*; *large language models*.

ABSTRACT

Under normal communication conditions, linguistic elements are articulated into a cohesive set, integrated with extralinguistic information, forming what is known as discourse. Coreference, alongside anaphora – which it often accompanies –, is a phenomenon inextricable from the construction of this structure, permeating all linguistic communication. Therefore, in the context of Natural Language Processing, coreference resolution is essential for understanding discourse, contributing to the access to this complex network of connections underlying it. This is a challenging task that relies on both linguistic and extralinguistic skills – such as comprehension of textual context, and world knowledge – and facilitates the execution of various other natural language processing tasks. Currently, most of the best-performing coreference resolution models rely on supervised machine learning. However, the performance of this approach is heavily dependent on large annotated corpora, a costly and scarce resource in Portuguese. Given this scenario, the use of prompt engineering on pre-trained large language models emerges as a potentially less expensive alternative to models that use supervised learning, as it does not require annotated data for training. This study evaluates the performance of this strategy for entity coreference resolution in Portuguese by two language models: ChatGPT-4o and a version of ChatGPT-4 customized with the incorporation of a system prompt and knowledge files. The models were tested using two series of zero and few-shot prompts. Furthermore, a segment of the Summ-it++ corpus – the second largest corpus with coreference information in Portuguese – was revised based on the OntoNotes guidelines, resulting in a revised version that was used alongside the original version as test corpora. Both corpora were harmonized according to the CorefUD format, allowing the evaluation of the results using the CorefUD scorer, which encompasses the main metrics in the field. In the CoNLL Score, the tests reached a performance 17.75% lower than the state-of-the-art model on the Summ-it++ corpus was reached. While it did not achieve state-of-the-art levels, the approach proved promising, delivering performance that, although inferior, was nonetheless relevant compared to the best models for the same corpus, suggesting potential for future improvements.

Keywords: coreference resolution; prompt engineering; large language models.

LISTA DE FIGURAS

Figura 1 – Modelo de discurso 18
Figura 2 – Ilustração da relação anafórica entre as expressões “Marcio” e “Ele”, na frase do exemplo *a* 19
Figura 3 – Ilustração da relação de catáfora entre as expressões “seu” e “Marina”, na frase do exemplo *c* 20
Figura 4 – Ilustração do caso de anáfora com antecedentes coordenados do exemplo *e* 21
Figura 5 – Ilustração das relações de correferência no exemplo *f* 22
Figura 6 – Ilustração das relações de correferência com e sem anáfora no exemplo *l* 24
Figura 7 – Ilustração da relação de anáfora sem correferência no exemplo *m* 25
Figura 8 – Segmento do corpus Coref-PT 61
Figura 9 – Exemplo de trecho do corpus Summ-it++ na formatação CoNLL-U com informações de correferência na formatação CorefUD 63
Figura 10 – Segmento do corpus Summ-it++ 66
Figura 11 – Segmento do corpus Summ-it++ 66
Figura 12 – Segmento do corpus Summ-it++ 67
Figura 13 – Segmento do corpus Summ-it++ 67
Figura 14 – Segmento do corpus Summ-it++ 68
Figura 15 – Segmento do corpus Summ-it++ 68
Figura 16 – Segmento do corpus Summ-it++ 69
Figura 17 – Segmento do corpus Summ-it++ 69
Figura 18 – Etapa de verificação da corrente de *prompts* 79
Figura 19 – Esquema de formatação de resposta do modelo 85
Figura 20 – Gráfico comparativo da média de *recall* dos modelos e abordagens testados 88
Figura 21 – Gráfico comparativo da média de precisão dos modelos e abordagens testados 89

LISTA DE TABELAS

Tabela 1 – Exemplos de corpora relevantes do inglês e seus respectivos enfoques e tamanhos 30
Tabela 2 – Comparação entre os 3 mais relevantes corpus do português anotados com informações de correferência 33
Tabela 3 – Resultados dos modelos propostos por Fonseca *et al.* (2016) 39
Tabela 4 – Resultados obtidos pelo modelo proposto por Fonseca *et al.* (2017) 40
Tabela 5 – Resultados obtidos pelo modelo proposto por Rocha *et al.* (2017) 41
Tabela 6 – Resultados obtidos pelo modelo proposto por Fonseca *et al.* (2018) 43
Tabela 7 – Resultados obtidos por Lima *et al.* (2018) 44
Tabela 8 – Resultados obtidos por Cruz *et al.* (2018) na resolução do corpus Corref-PT 45
Tabela 9 – Resultados obtidos por Yang *et al.* (2022) 49
Tabela 10 – Resultados obtidos por Gan *et al.* (2024) em avaliação automática 50
Tabela 11 – Resultados obtidos por Gan *et al.* (2024) em avaliação manual 51
Tabela 12 – Resultados obtidos por Le *et al.* (2023) em testes com menções preditas 52
Tabela 13 – Resultados obtidos por Le *et al.* (2023) em testes com menções douradas 53
Tabela 14 – Resultados do InstructGPT em corpora multilíngues 53
Tabela 15 – Resultados obtidos por Hicke *et al.* (2024) 54
Tabela 16 – Escopo dos principais corpora anotados com informação de correferência 69
Tabela 17 – Comparação dos resultados obtidos pelos modelos deste trabalho aos dos resultados de outros modelos de RC do português 90
Tabela 18 – Comparação dos resultados do presente trabalho com os de Le *et al*. (2023) 91
Tabela 19 - Distribuição da classe gramatical dos núcleos de menções na resolução do corpus original pelos modelos testados 92
Tabela 20 – Distribuição da classe gramatical dos núcleos de menções na resolução do corpus revisado pelos modelos testados 93

LISTA DE ABREVIATURAS E SIGLAS

IA Inteligência Artificial

PLN Processamento de Língua Natural

LLM *Large Language Models*

RC Resolução de Correferência

RNR Rede Neural Recorrente

SUMÁRIO

1 INTRODUÇÃO 10
1.1 OBJETIVOS 13
1.1.1 Objetivo geral 13
1.1.2 Objetivos específicos 13
1.2 JUSTIFICATIVA 14
1.3 LIMITAÇÕES 14
1.4 ORGANIZAÇÃO DO DOCUMENTO 16
2 FUNDAMENTAÇÃO TEÓRICA 17
2.1 ANÁFORA, CATÁFORA E CORREFERÊNCIA 17
2.2 CLASSIFICAÇÃO DO PROBLEMA DE RESOLUÇÃO DE CORREFERÊNCIA 27
2.3 CORPORA 28
2.3.1 Corpora do português 32
2.4 MODELOS DE RESOLUÇÃO DE CORREFERÊNCIA 34
2.4.1 Modelos baseados em características 35
2.4.2 Modelos baseados em redes neurais recorrentes 36
2.4.3 Modelos baseados em conhecimento 36
2.4.4 Modelos baseados em *transformers* 37
2.5 TRABALHOS RELACIONADOS 38
2.5.1 Modelos de resolução de correferência em português 38
2.5.2 Uso de engenharia de *prompt* para a resolução de correferência 46
2.6 MÉTRICAS E *SCORERS* 55
2.6.1 F^1 56
2.6.2 MUC 56
2.6.3 B^3 (B-Cubed) 57
2.6.4 CEAF 58
2.6.5 BLANC 58
2.6.6 LEA 59
2.6.7 Combinação de métricas e *scorers* 60
3 METODOLOGIA 61

3.1 CORPUS DE TESTAGEM 61
3.1.1 Escolha do corpus 61
3.1.2 CorefUD 62
3.1.3 Harmonização do corpus 64
3.1.3.1 Revisão de anotações de correferência 65
3.1.3.2 Diretrizes do OntoNotes 71
3.2 ELABORAÇÃO DOS *PROMPTS* 73
3.2.1 Abordagem A 81
3.2.2 Abordagem B 81
3.3 MODELOS 82
3.3.1 ChatGPT-4o 82
3.3.2 GPT customizado 83
3.4 CONVERSÃO DAS RESPOSTAS PARA O FORMATO COREFUD 84
3.5 PONTUAÇÃO 85
4 RESULTADOS 88
5 CONCLUSÃO 94
REFERÊNCIAS 97
APÊNDICE A – ESTRUTURA DE *PROMPTS* ABORDAGEM A 101
APÊNDICE B – ESTRUTURA DE *PROMPTS* ABORDAGEM B 102
APÊNDICE C – PONTUAÇÃO DOS MODELOS TESTADOS 103

1 INTRODUÇÃO

Ao interpretar um discurso, cria-se um modelo mental - denominado **modelo de discurso** - que contém a representação das entidades mencionadas no mesmo, bem como as relações entre elas e suas propriedades (Karttunen, 1969, *apud* Jurafsky *et al.*, 2024). As expressões linguísticas que se referem a essas entidades são suas **menções**, e as entidades às quais se referem são seus **referentes** (Jurafsky *et al.*, 2024). Quando mais de uma menção se refere à mesma entidade do discurso, se dá o fenômeno de correferência (Jurafsky *et al.*, 2024).

a) "[Márcia]$_i$ atende os pacientes muito bem. Todos eles adoram [a chefe do departamento de nutrição]$_i$."

De acordo com Mitkov (1999), para que duas expressões possam ser correferentes, elas precisam, além de outros critérios, concordar em número e gênero e se alinhar em termos de suas propriedades semânticas – os significados e papéis que elas transmitem. Em *a*, as expressões "Márcia" e "a chefe do departamento de nutrição" se referem à mesma entidade no modelo de discurso - a pessoa Márcia - e têm, portanto, relação de correferência. Nesse exemplo, descrição fornecida por "a chefe do departamento de nutrição" delimita a referência a uma pessoa que cumpra esse papel específico. Dentro do contexto, Márcia é a única entidade mencionada que satisfaz esses critérios (gênero feminino, singular, profissional da área da saúde, etc.) – o que indica uma relação de correferência. Outros critérios, como proximidade das expressões e saliência no discurso, ajudam a indicar uma possível relação de correferência em casos ambíguos (Mitkov, 1999).

A resolução de correferência (RC) é a tarefa de identificar e agrupar as menções que se referem à mesma entidade no modelo de discurso (Jurafsky *et al.*, 2024). Um bom desempenho na resolução de correferência pode tornar a informação em um texto mais acessível e organizada para sistemas de Processamento de Língua Natural[1] (PLN) e outras tarefas - o que torna essa tarefa importante para várias das aplicações do PLN, como a sumarização, vinculação

[1] Para este trabalho, optou-se pelo termo "Processamento de Língua Natural" em lugar do termo mais comumente utilizado, "Processamento de Linguagem Natural". Essa opção se dá por conta da distinção amplamente aceita na área da Linguística entre linguagem - competência/faculdade física, fisiológica e psíquica - e língua - produto social dessa faculdade da linguagem, conjunto de signos e convenções que permite seu exercício (Saussure, 2006) - e pelo entendimento de que o objeto sobre o qual se dão as ações de PLN é, de fato, a língua, como manifestação concreta e observável da linguagem.

de entidades, reconhecimento de entidades nomeadas, resposta a perguntas e análise de sentimentos (Liu *et al.*, 2023).

A relação de correferência também pode se dar entre menções a eventos (Jurafsky *et al.*, 2024; Sukthanker *et al.*, 2020).

b) “[Aluá conseguiu ingressar em um ótimo programa de mestrado]$_i$. [A conquista]$_i$ deixou todos muito orgulhosos.”

No exemplo *b*, as expressões “Aluá conseguiu ingressar em um ótimo programa de mestrado” e “a conquista” são menções que se referem ao mesmo evento no modelo de discurso: o fato de Aluá ter ingressado em um dado programa de mestrado. Trata-se, portanto, de um caso de correferência de eventos. Uma vez identificadas as menções, as mesmas estratégias de RC utilizadas para correferência de entidades costumam ser aplicadas para a resolução de correferência de eventos (Jurafsky *et al.*, 2024).

A RC de eventos, contudo, é consideravelmente mais desafiadora do que a de entidades. Determinar se duas menções em um texto se referem ao mesmo evento é uma tarefa particularmente complexa, que envolve a análise da ontologia dos eventos, da estrutura lógica de uma sentença e dos papéis semânticos de seus componentes. Diferentemente da constatação de fatos, eventos são relacionados de diferentes maneiras à fatores tempo, modo e modalidade. Além disso, a perspectiva por meio da qual um evento é apresentado também desempenha papel crucial nessa determinação, pois a forma como é descrito pode influenciar sua interpretação, atribuindo diferentes papéis aos participantes e dificultando a determinação de identidade em relação a outros eventos. Um exemplo típico envolve os conceitos de "compra" e "venda", onde a mesma transação pode ser descrita de formas distintas, dependendo de qual perspectiva — do agente ou do paciente — é enfatizada.

Dessa forma, definir a correferência entre menções a esse tipo de evento é um processo que, inevitavelmente, engloba a análise de muitas dimensões e nuances. As descrições linguísticas nem sempre são simétricas ou plenamente intercambiáveis, e inferências incorretas podem surgir quando essas nuances são desconsideradas (Williams, 2021). Assim, uma vez que essa tarefa depende de uma série de componentes de extração de informação cujos resultados são, ainda, muito ruidosos (Jurafsky *et al.*, 2024; Sukthanker *et al.*, 2020), a maior parte dos modelos e corpora da área de RC não envolve RC de eventos (Liu *et al.*, 2023).

Nos últimos anos, o foco da pesquisa em resolução de correferência passou de modelos baseados em regras e em aprendizado de máquina tradicional para abordagens envolvendo aprendizado de máquina profundo que utilizam informações contextuais e dados explícitos de bases de conhecimento (Liu *et al.*, 2023). Atualmente, a maioria dos modelos estado da arte em RC nos corpora mais relevantes da área são baseados em grandes modelos de linguagem (LLMs) pré-treinados que utilizam a arquitetura de *transformers* (Liu *et al.*, 2023).

Estes modelos, contudo, dependem amplamente de aprendizado supervisionado[2], que requer grandes bases de dados anotadas – visto que a qualidade do aprendizado depende, principalmente, do tamanho dessas bases de dados – e é altamente oneroso (Yang *et al.*, 2022). Além disso, as bases de dados anotadas com informações de correferência disponíveis são consideravelmente menores do que as bases de dados utilizadas para tarefas mais gerais de PLN, como anotação morfossintática (Yang *et al.*, 2022) - especialmente em línguas como o português (Fonseca *et al.*, 2018), o que impacta a eficácia do aprendizado supervisionado para essa tarefa.

Recentemente, observa-se o surgimento de uma tendência ao uso de engenharia de *prompts* com LLMs pré-treinados como estratégia para lidar com tarefas de PLN com poucos recursos (Yang *et al.*, 2022). Diferentemente dos ajustes finos (Brown *et al.*, 2020; Wei *et al.*, 2022 *apud* Yang *et al.*, 2022), que requerem a atualização dos pesos do modelo pré-treinado a partir do *input* de um grande volume de dados anotados, essa estratégia dispensa aprendizado supervisionado e mantém os pesos inalterados (Yang *et al.*, 2022). Em vez de treinar um modelo para tarefas específicas, essa técnica adapta essas tarefas para que correspondam à tarefa original dos LLMs por meio de engenharia de *prompt* (Liu *et al.*, 2021 *apud* Yang *et al.*, 2022).

Em novembro de 2023, a OpenAI introduziu uma ferramenta de criação de versões customizadas do ChatGPT, denominadas GPTs (OpenAI, 2024c). Baseados no modelo ChatGPT-4, os GPTs permitem a combinação de diferentes componentes, denominados *GPT internals*, que incluem *system prompts*, arquivos de conhecimento e funcionalidades externas (Liu *et al*., 2024; Zhang *et al*., 2024).

Os *system prompts* são, de acordo com Xu *et al*. (2024) instruções que funcionam como um "prefixo padrão implicitamente adicionado ao *input* do usuário" durante a fase de inferência dos LLMs. Esses *prompts* operam de forma subjacente às interações com o usuário e contêm diretrizes detalhadas que ajudam o modelo a interpretar melhor o contexto e fornecer

[2] Abordagem de aprendizado de máquina na qual um modelo é treinado com exemplos que consistem em uma entrada e a respectiva saída desejada (Pusteyovsky *et al.*, 2013).

respostas mais adequadas às solicitações recebidas. Pape *et al*. (2024) apontam que o uso de s*ystem prompts* com instruções detalhadas pode transformar LLMs genéricas em ferramentas específicas com uma demanda de recursos mínima e sem a necessidade de ajustes finos com aprendizado supervisionado.

Ainda há uma escassez de estudos sobre o uso de *prompting* em LLMs para a tarefa de resolução de correferência (Sanh *et al.*, 2021 *apud* Yang *et al.*, 2022). Diante disso, o presente trabalho tem como objetivo avaliar o desempenho do uso dessa abordagem para a RC de entidades em português em dois modelos, como potencial estratégia para contornar a escassez de recursos anotados nessa área em português.

1.1 OBJETIVOS

1.1.1 Objetivo geral

O objetivo geral deste trabalho é avaliar o desempenho do uso da engenharia de *prompt* no ChatGPT-4o (OpenAI, 2024b) e em uma versão customizada do ChatGPT-4 (OpenAI, 2024a; OpenAI, 2024c) para tarefa de resolução de correferência de entidade na língua portuguesa.

1.1.2 Objetivos específicos

Os objetivos específicos deste trabalho são:

1) Pesquisar e documentar o estado da arte em resolução de correferência no português, bem como o cenário atual do uso da engenharia de *prompt* em LLMs na tarefa de resolução de correferência.
2) Desenvolver *prompts* para a adaptação da tarefa de resolução de correferência à tarefa original de *chatbots* baseados em LLMs.
3) Desenvolver um GPT customizado para a tarefa de resolução de correferência.
4) Implementar a conversão das respostas geradas pelos modelos testados para o formato de anotação compatível com os corpora de teste utilizados, e avaliar os modelos por meio das métricas mais frequentemente utilizadas na área.
5) Analisar os resultados obtidos levando em consideração os resultados dos demais modelos de RC em português.

1.2 JUSTIFICATIVA

A correferência, junto da anáfora – com a qual frequentemente ocorre –, desempenha um papel essencial na construção e compreensão do discurso, se tratando ele de uma estrutura coesa, conectada a um modelo mental representativo (Mitkov, 2002; Karttunen, 1969 *apud* Jurafsky *et al.*, 2024). Uma boa capacidade de resolução desse importante fenômeno contribui para o desempenho de sistemas de Processamento de Língua Natural em diversas outras tarefas (Liu *et al.*, 2023).

Os modelos estado da arte para resolução de correferência em inglês são predominantemente baseados em aprendizado supervisionado (Liu *et al.*, 2023). Devido à escassez de recursos disponíveis para esse tipo de abordagem no português, os modelos estado da arte utilizados na resolução dos principais corpora anotados com informação de correferência do idioma ainda são baseados em regras (Fonseca *et al.*, 2018). Essa abordagem, embora funcional, já foi superada em línguas com mais recursos, como o inglês (Liu *et al.*, 2023).

Nos últimos anos, tem-se observado uma tendência crescente ao uso de engenharia de *prompts* em LLMs pré-treinados para lidar com tarefas de PLN que dispõem de poucos recursos (Yang *et al.*, 2022). Essa abordagem tem sido particularmente útil em contextos onde a disponibilidade de corpora anotados é limitada, por dispensar a necessidade de aprendizado supervisionado (Liu *et al.*, 2021 apud Yang *et al.*, 2022).

Diante do cenário ainda incipiente da resolução de correferência no português, verifica-se a necessidade de explorar alternativas capazes de contornar a escassez de recursos disponíveis e promover avanços na área. Dessa forma, é proposto um sistema de RC no português baseado em engenharia de *prompt* utilizando LLMs pré-treinados, explorando também o uso de *system prompts* para a customização de um modelo, visando avaliar a viabilidade dessa abordagem — ainda não explorada para a tarefa de RC nesse idioma — como uma solução promissora para mitigar essas limitações atuais.

1.3 LIMITAÇÕES

Conforme observado por Jurafsky *et al.* (2024) e Sukthanker *et al.* (2020), a RC de eventos apresenta caráter significativamente mais desafiador quando comparada à RC de entidades. O desenvolvimento de modelos eficazes nessa tarefa é mais complexo, devido a,

entre outros fatores, uma maior dificuldade nas etapas de extração de informações e identificação de menções (Jurafsky *et al.*, 2024; Sukthanker *et al.*, 2020). Além disso, a avaliação desses modelos é comprometida pela frequente inconsistência nas regras de anotação de correferência nos diferente corpora disponíveis - ainda mais acentuada na anotação de correferência de eventos quando comparada à de correferência de entidades - o que afeta a confiabilidade e uniformidade dos resultados (Mitkov, 2022). Não por acaso, a maioria dos estudos e modelos desenvolvidos na área de RC sequer aborda a correferência de eventos (Liu *et al.*, 2023). Diante desse cenário, optou-se pela não inclusão da resolução de correferência de eventos no presente estudo. Também não é incluída no estudo a resolução de casos de anáfora indireta ou com antecedentes coordenados. Essa escolha foi motivada pela complexidade envolvida na anotação e avaliação desses fenômenos no formato utilizado[3]. Dessa forma, o presente trabalho se concentra exclusivamente na resolução de correferência de entidades, englobando a resolução de anáforas somente quando estas coocorrem com as instâncias de correferência e têm somente um antecedente.

Ademais, devido a limitações de recursos, este trabalho foca exclusivamente na avaliação do desempenho do ChatGPT-4 e de seu sucessor, o ChatGPT-4o, sem explorar suas versões anteriores ou outros grandes modelos de linguagem, como BERT (Devlin *et al.*, 2019 *apud* Yang *et al.*, 2022), e LLAMA (Touvron *et al.*, 2023 *apud* Gan *et al.*, 2024). A escolha se justifica pelo fato de que o ChatGPT-4 obteve o melhor desempenho dentre os modelos testados até então para o uso da engenharia de *prompt* na tarefa tradicional de RC (Gan *et al.*, 2024). Assim, optou-se por explorar uma versão customizada do ChatGPT4 e o ChatGPT-4o, partindo da suposição de que ele possa oferecer resultados superiores aos de seu antecessor e considerando as melhorias significativas demonstradas pelo modelo no tratamento de textos em idiomas que não são o inglês (OpenAI, 2024b). Além disso, pretendia-se, inicialmente, incluir no presente estudo o uso de LLMs voltados ao português. Foram feitos testes preliminares utilizando modelos da família Sabiá, família de LLMs treinados em textos em português que tem alcançado resultados promissores, obtendo desempenho superior ao do ChatGPT-4 e ChatGPT-3.5 na resolução de provas em português (Almeida *et al.*, 2024). No entanto, os modelos testados se mostraram incapazes de produzir respostas no formato escolhido para o trabalho.

[3] O formato CorefUD, descrito na seção 3.1.2.

Por último, foram testadas apenas duas combinação de *prompts*. Como apontado por Oppenlaender *et al.* (2023 *apud* Knoth *et al.*, 2024, tradução nossa), "a criação de *prompts* eficientes [...] é um desafio, uma vez que exige um extenso processo de tentativa e erro, e uma avaliação rigorosa de diferentes estratégias de *prompt* em pares de *input* e *output* e grandes corpora". Assim, também devido a limitações de tempo e recursos, e dado o caráter extremamente oneroso do processo de criação de *prompts*, optou-se por utilizar as combinações de *prompts* que apresentaram os melhores resultados dentre uma pequena quantidade de opções testadas manualmente.

1.4 ORGANIZAÇÃO DO DOCUMENTO

O trabalho seguirá a seguinte trajetória: no Capítulo 2, na seção 2.1 e 2.2, é apresentada uma introdução aos conceitos de anáfora, catáfora e correferência. A seção 2.2 oferece uma breve classificação do problema de resolução de correferência. A seção 2.3, por sua vez, apresenta alguns dos principais corpora anotados com informação de correferência, tanto do inglês como do português e de outras línguas. Na seção 2.4, é feita uma breve revisão das abordagens e métodos utilizados para a tarefa de resolução de correferência até o momento. A seção 2.5 apresenta um panorama dos trabalhos realizados até o momento sobre o uso da engenharia de *prompt* em LLMs na resolução de correferências, e a seção 2.6 expõe as métricas mais comumente utilizadas na avaliação de modelos de resolução de correferência. No terceiro capítulo, é apresentada a metodologia utilizada na pesquisa. Na seção 3.1, são descritos os procedimentos adotados para a escolha e ajuste do corpus de testagem. Na seção 3.2, detalham-se os procedimentos para a criação dos *prompts* utilizados. A seção 3.3 detalha os modelos avaliados, incluindo o processo de configuração do GPT customizado. As seções 3.4 e 3.5 abordam os métodos para a aplicação dos *prompts*, bem como para a conversão e pontuação das respostas obtidas pelo modelo. No quarto capítulo, são discutidos os resultados obtidos. Finalmente, no capítulo 5, são apresentadas as conclusões da pesquisa.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 ANÁFORA, CATÁFORA E CORREFERÊNCIA

Um **modelo de discurso** é um modelo mental que uma pessoa (ou um sistema de Processamento de Língua Natural) constrói incrementalmente ao interpretar um texto, que contém representações das entidades mencionadas nesse texto, assim como das propriedades dessas entidades e as relações entre elas (Karttunen, 1969, *apud* Jurafsky *et al.*, 2024). Quando uma expressão linguística se refere a uma entidade do modelo de discurso, tal expressão é uma **menção** (ou expressão referente), e a entidade à qual se refere é seu **referente** (Jurafsky *et al.*, 2024). De acordo com Jurafsky *et al.* (2024), quando uma entidade é mencionada pela primeira vez em um discurso, uma representação dela é **evocada** no modelo desse discurso. Nas suas demais menções, essa representação é **acessada** a partir do modelo (Jurafsky *et al.*, 2024).

A Figura 1 representa esse processo de construção de um modelo de discurso a partir do enunciado "Márcia atende os pacientes muito bem. Todos eles adoram a chefe do departamento de nutrição".

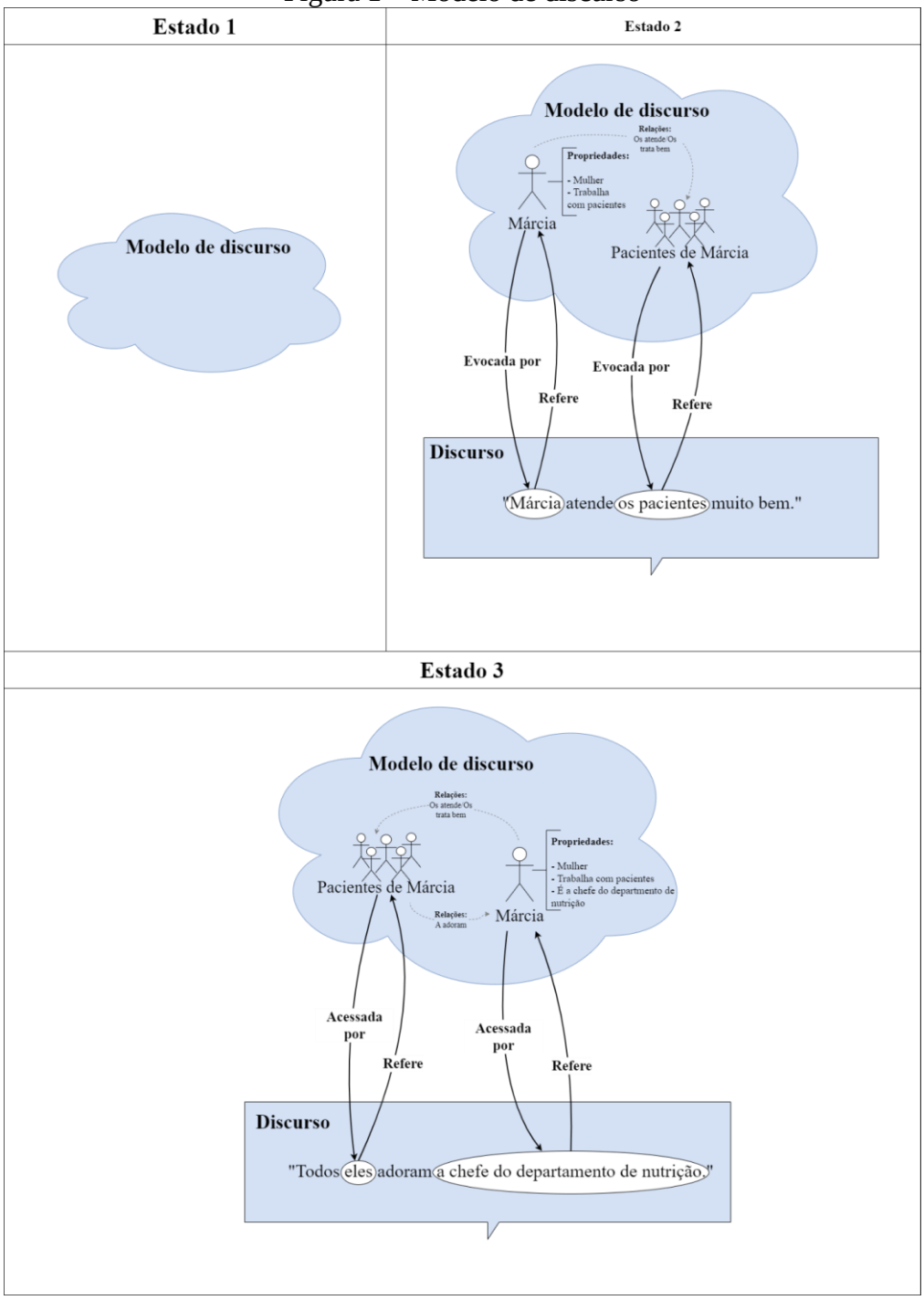
O modelo começa vazio. A expressão "Márcia" é uma *menção* que tem por *referente* a pessoa Márcia - entidade ainda não mencionada no discurso. Dessa forma, essa menção *evoca* uma representação da entidade Márcia no modelo de discurso. Em seguida, a menção "os pacientes" evoca no modelo uma representação do seu referente, o grupo dos pacientes da pessoa Márcia.

Na frase seguinte, a expressão "eles" é uma nova menção ao grupo dos pacientes de Márcia, entidade já evocada anteriormente. Dessa forma, essa menção não evoca, mas *acessa* a representação já existente dessa entidade no modelo de discurso. O mesmo ocorre com a expressão "a chefe do departamento de nutrição", que acessa a representação da entidade Márcia, previamente evocada.

Ao longo desse processo, informações referentes a essas entidades - as relações entre elas e suas propriedades - são incrementalmente adicionadas ao modelo. A partir da primeira frase, é possível inferir algumas propriedades da entidade Márcia: trata-se de uma mulher que trabalha, de alguma maneira, com pacientes. Também sabe-se que Márcia trata bem os seus pacientes, o que revela uma relação entre as entidades representadas no discurso. Na frase seguinte, é revelada uma nova informação sobre a relação entre as duas entidades - a de que os

pacientes de Márcia a adoram - e sobre as propriedades da entidade Márcia - a de que ela é a chefe do departamento de nutrição.

Figura 1 – Modelo de discurso



Fonte: autora

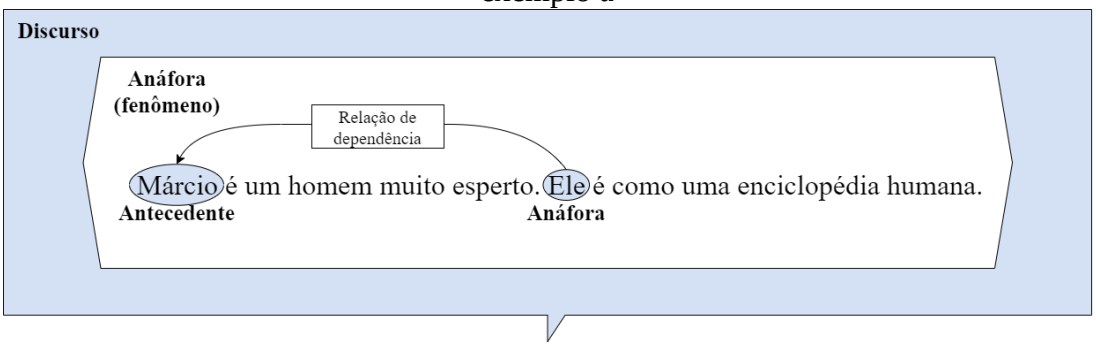
Em condições normais, a comunicação linguística transcende a mera enunciação de itens isolados. Se trata não de um conjunto de declarações avulsas e desconexas, mas sim de

um fluxo integrado de ideias (Mitkov, 2002). Essa interdependência entre os elementos presentes nas partes de um discurso, de acordo com Mitkov (2002), é explicável pelo fenômeno da **coesão**. Este fenômeno, essencial para a construção do sentido, fundamenta o processo pelo qual a interpretação de um componente do discurso está ligada à de outros (Mitkov, 2002).

A **anáfora** é descrita por Halliday *et al.* (1976 *apud* Mitkov 2002) como um fenômeno de coesão que aponta para uma expressão linguística previamente mencionada no discurso. Conforme Mitkov (2002), a essa expressão que "aponta para um item anterior" se dá o nome de **anáfora**[4], e a expressão para o qual ela aponta é o seu **antecedente**.

a) "[Márcio]$_i$ é um homem muito esperto. [Ele]$_i$ é como uma enciclopédia humana."
b) "[Nica]$_i$ era um cão muito carinhoso, e também extremamente inteligente. [O animal]$_i$ surpreendia a todos com sua facilidade para aprender truques."

Figura 2 – Ilustração da relação anafórica entre as expressões "Marcio" e "Ele", na frase do exemplo *a*



Fonte: autora

Conforme ilustrado na Figura 2, no exemplo *a*, a expressão "ele" depende da expressão "Márcio" para ser compreendida. Desta forma, "ele" trata-se de uma anáfora que aponta para "Márcio", seu antecedente. Analogamente, no exemplo *b*, a expressão "o animal", uma anáfora, aponta para a expressão "Nica", seu referente.

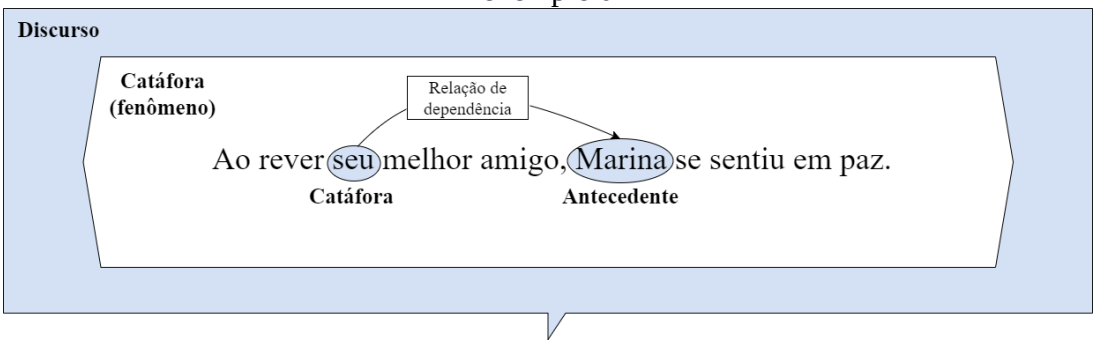
A **catáfora**, por sua vez, é um fenômeno que opera de maneira análoga à anáfora, mas com a ordem das expressões invertida (Mitkov, 2002). Em uma relação catafórica, uma expressão aponta para um item que o segue, como no exemplo a seguir:

[4] No inglês, se faz a distinção entre *anaphora*, anáfora como fenômeno de coesão, e *anaphor*, a expressão que se refere a um antecedente (Mitkov, 2002). Em português, ambos os conceitos são chamados de anáfora.

c) “Ao rever [seu]$_i$ melhor amigo, [Marina]$_i$ se sentiu em paz.”

No exemplo *c*, conforme ilustrado na Figura 3, a expressão “seu” vem antes do item para o qual aponta, “Marina”. Dessa forma, “seu” é uma catáfora que aponta para “Marina”, seu antecedente.

Figura 3 – Ilustração da relação de catáfora entre as expressões “seu” e “Marina”, na frase do exemplo *c*



Fonte: autora

Mitkov (2002) propõe algumas variedades de anáfora[5] baseando-se em seus aspectos formais. Entre elas, destaca-se a anáfora nominal, que ocorre quando um pronome ou sintagma nominal definido tem por antecedente um sintagma nominal não pronominal. Este tipo de relação é a mais amplamente estudada na área do PLN até o momento (Mitkov, 2002).

Além disso, a anáfora pode se referir a sintagmas verbais, verbos ou advérbios locativos e temporais, caracterizando-se, nesse caso, como anáfora verbal (exemplos *g* e *m*[6]) e adverbial (Mitkov, 2002). Outra variedade mencionada por Mitkov (2002) é a anáfora zero ou elipse, em que a forma anafórica é omitida, podendo ter por antecedente pronomes, substantivos e verbos ou frases verbais, como ocorre no exemplo.

d) “[Leônidas]$_i$ era muito egoísta. [Ø]$_i$ Só fazia o que [Ø]$_i$ queria, e quando [Ø]$_i$ queria.”

No exemplo *d*, é utilizada a elipse (marcada no exemplo com o símbolo Ø) no lugar da repetição do antecedente “Leônidas” para recuperá-lo como agente de “fazia” e “queria”, tratando-se, portanto, de um caso de anáfora zero.

---

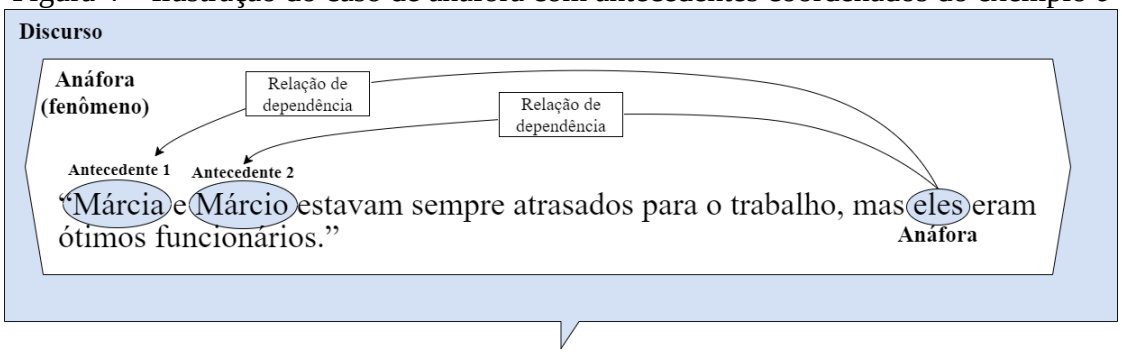
[5] Serão expostos apenas alguns dos tipos de anáfora elencados por Mitkov (2002). Outros tipos podem ser consultados em sua obra.

[6] Páginas 22 e 24, respectivamente.

Uma anáfora também pode ter **antecedentes coordenados** (Mitkov, 2002), como no exemplo *e*.

e) "[Márcia]$_i$ e [Márcio]$_a$ estavam sempre atrasados para o trabalho, mas [eles]$_{ia}$ e eram ótimos funcionários."

Figura 4 – Ilustração do caso de anáfora com antecedentes coordenados do exemplo *e*



Fonte: autora

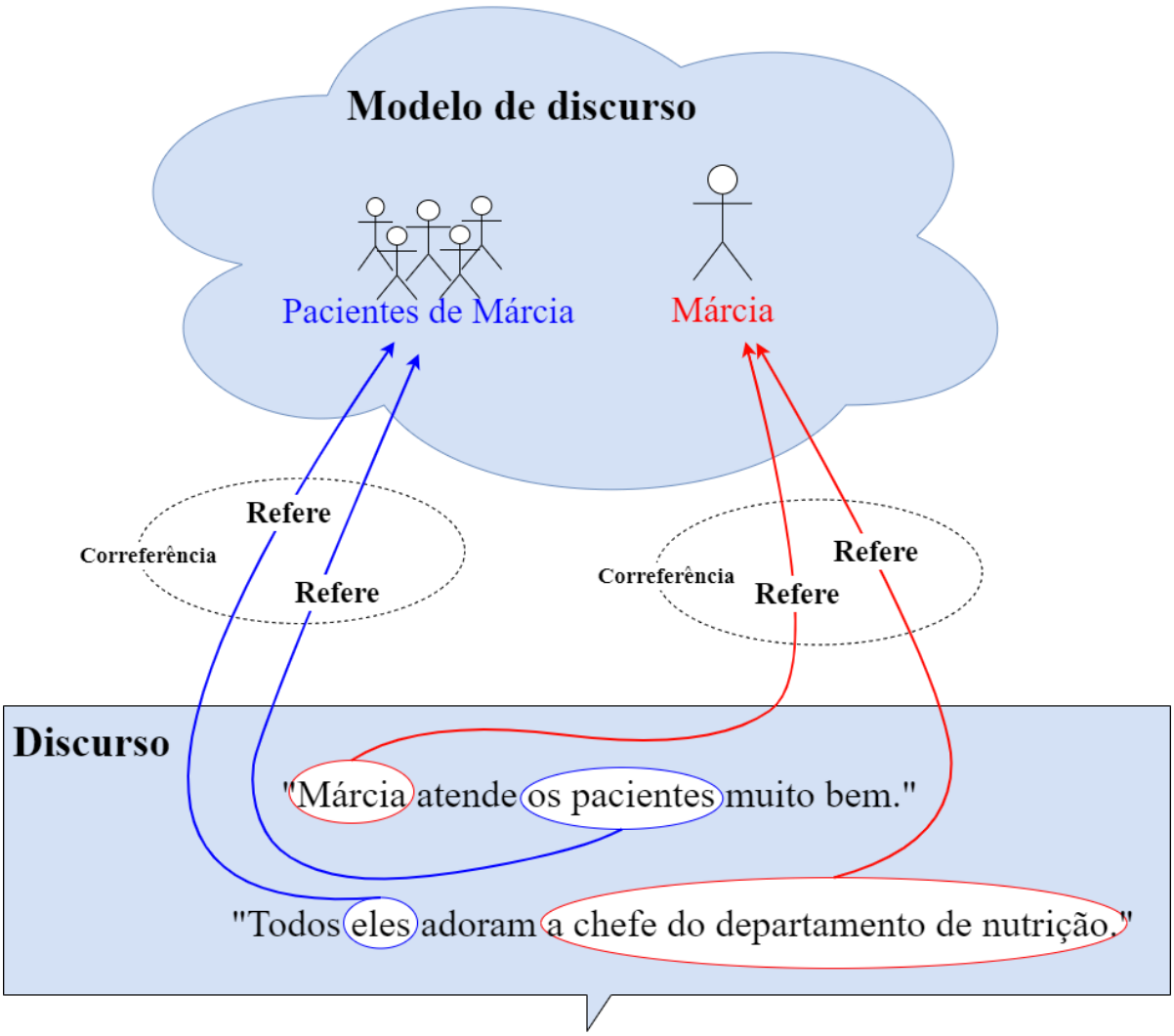
Neste tipo de anáfora, um elemento anafórico tem mais de um antecedente. Conforme ilustrado na Figura 4, a expressão "eles", no exemplo *e*, tem por antecedentes as expressões "Márcia" e "Márcio", simultaneamente.

A **correferência**, por sua vez, de acordo com Jurafsky *et al.* (2024), é um fenômeno que diz respeito à relação entre os elementos linguísticos no que concerne a seus referentes em um dado modelo de discurso. Quando dois ou mais elementos têm o mesmo referente em um modelo do discurso, estes elementos são **correferentes** (Jurafsky *et al.*, 2024). Uma entidade que possui apenas uma menção em um texto é chamada de ***singleton*** (Jurafsky *et. al*, 2024).

f) "[Márcia]$_i$ trata [os pacientes]$_a$ muito bem. [Todos eles]$_a$ adoram [a chefe do departamento de nutrição]$_i$."

No exemplo *f*, previamente ilustrado pela Figura 1, as expressões "Márcia" e "a chefe do departamento de nutrição" se referem à mesma entidade no modelo de discurso, a pessoa Márcia, e são, portanto, menções correferentes. A mesma relação se dá entre as expressões "os pacientes" e "eles", que se referem ao grupo composto pelos pacientes de Márcia. Essas relações de correferência são ilustradas na Figura 5.

Figura 5 – Ilustração das relações de correferência no exemplo *f*



Fonte: autora

A relação de correferência pode, também, se dar entre menções a eventos (Jurafsky *et al.*, 2024):

g) "Rita se sentiu muito triste durante [o sepultamento de sua irmã]$_i$. [O ato de colocá-la em uma sepultura]$_i$ a parecia muito estranho."

No exemplo *g*, "o sepultamento de sua irmã" e "o ato de colocá-la em uma sepultura" se referem ao mesmo evento. São, portanto, menções a um evento com relação de correferência.

Frequentemente, relações de predicado e aposto se dão entre elementos que correferem (Mitkov, 2002), como "Márcio" e "o irmão gêmeo do Maurício", no exemplo *h*, e "Márcia" e "a irmã mais velha de Marília" no exemplo *i*.

h) “[Márcio]$_i$ é [o irmão gêmeo do Maurício]$_i$.”
i) “[Márcia]$_i$, [a irmã mais velha de Marília]$_i$, estava sempre pensando na próxima refeição.”

Relações de predicado e aposto que envolvam um conceito genérico ou um fato hipotético/uma possibilidade, contudo, costumam não ser consideradas correferenciais (Mitkov, 2002).

j) “[Vinícius]$_i$ agora era [um advogado]$_i$.”
k) “[Leônidas]$_i$, talvez [a pessoa mais incompetente da equipe]$_i$, era muito desagradável.”

No exemplo *j*, “um advogado” se trata de um conceito genérico que se aproxima mais de uma classificação do que de uma entidade. Esse tipo de conceito é, normalmente, considerado amplo demais para que seja parte de relações de correferência (Mitkov, 2002). Nesse caso, portanto, a expressão “um advogado” seria frequentemente considerada como não tendo relação de correferência com a menção “Vinícius”, mas sendo um atributo associado à entidade Vinicius. No exemplo *k*, por sua vez, a ideia de Leônidas ser a pessoa mais incompetente da equipe trata-se de uma hipótese, uma possibilidade, e não um fato concreto, de forma que “Leônidas” e “a pessoa mais incompetente da equipe” também não seriam normalmente considerados correferentes. No entanto, como aponta Mitkov (2002), a análise de casos como esses, bem como de outros relacionados à correferência, frequentemente depende de escolhas interpretativas e subjetivas do analista. Não há um consenso absoluto sobre a aplicação dessas relações, uma vez que a interpretação pode variar conforme o contexto e os critérios de cada pesquisador. Assim, a determinação de correferência envolve, inevitavelmente, algum grau de interpretação pessoal, devido à complexidade inerente da tarefa, como aponta Mitkov (2002).

Ao longo de um discurso, diversos elementos podem correferir. Essas várias menções a um mesmo evento ou entidade formam uma **cadeia de correferência** (Jurafsky *et al.*, 2024).

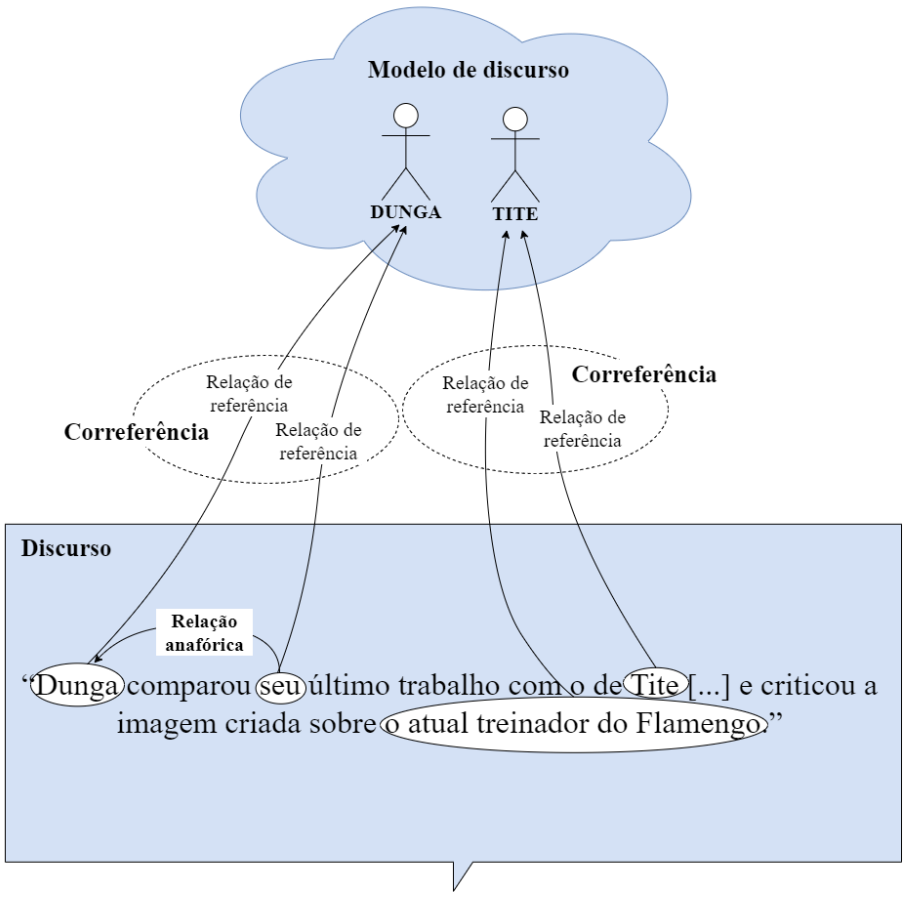
Os fenômenos de anáfora e correferência podem ou não coocorrer (Mitkov, 2002), como observa-se nos exemplos a seguir:

l) “Dunga comparou seu último trabalho com o de [Tite]$_i$ [...] e criticou a imagem criada sobre [o atual treinador do Flamengo]$_i$.”[7]

m) “Márcia [chegou atrasada ao trabalho naquele dia]$_i$, e Márcio também [o fez]$_i$.”

No exemplo *l*, ilustrado pela Figura 6, “seu” aponta para “Dunga” bem como ambos apontam para a mesma entidade no modelo de discurso, a pessoa Dunga. Essas expressões, portanto, têm, simultaneamente, relação anafórica e de correferência. Estes casos em que as anáforas e seus antecedentes se referem à mesma entidade no modelo de discurso - ou seja, correferem - Mitkov (2002) denomina **anáforas de identidade-de-referência** (*identity-of-reference*). Já, no mesmo exemplo, “Tite” e “o atual treinador do Flamengo” se referem à mesma entidade no modelo de discurso, mas os termos não dependem um do outro para sua interpretação - ou seja, são correferentes, mas não anafóricos.

Figura 6 – Ilustração das relações de correferência com e sem anáfora no exemplo *l*

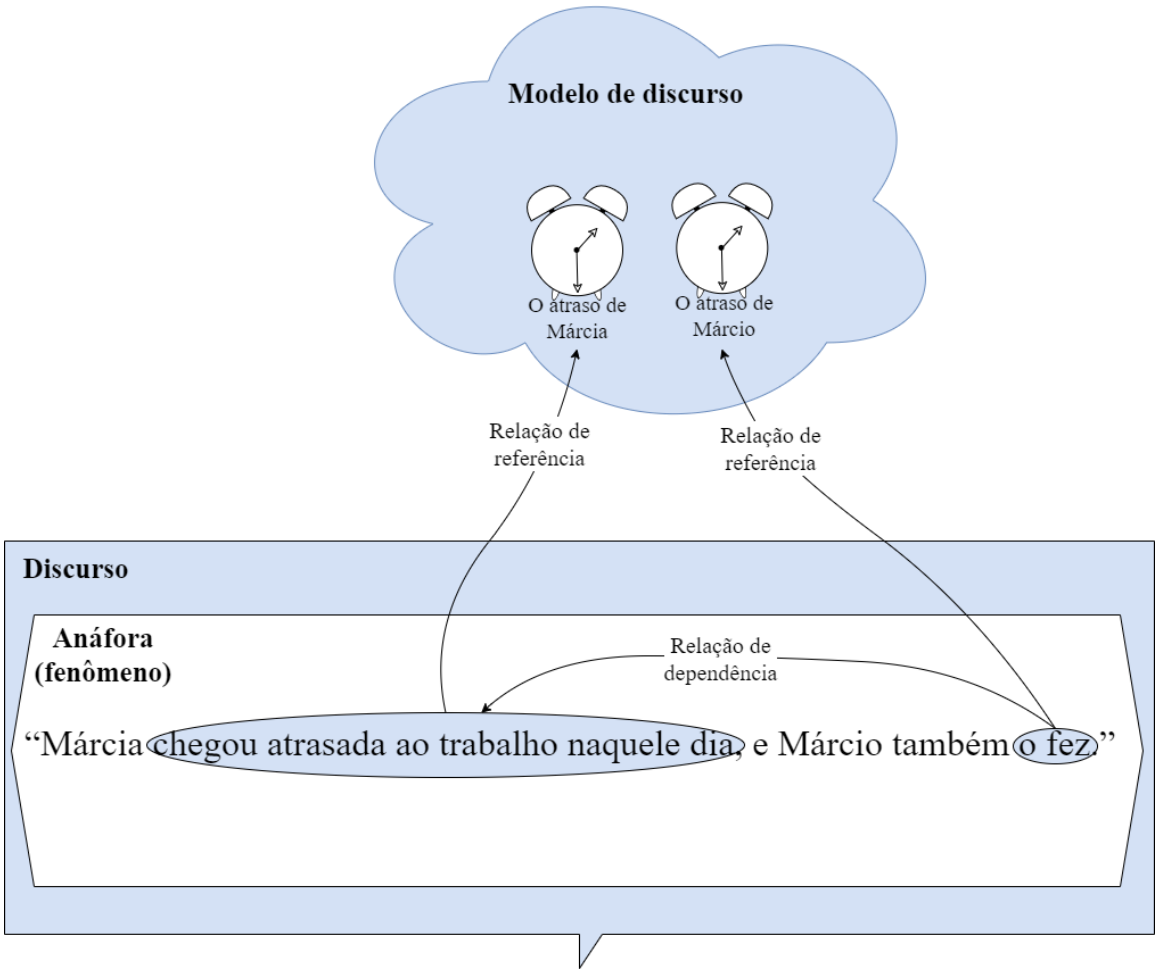


Fonte: autora.

[7] Fonte: https://www.cnnbrasil.com.br/esportes/futebol/dunga-detona-tite-na-selecao-nao-coloquei-minha-mae-para-dar-entrevista/

No exemplo *m*, ilustrado pela Figura 7, a expressão “o fez” aponta para “chegou atrasada no trabalho naquele dia”. No entanto, os eventos referidos - o atraso de Márcia e o atraso de Márcio - são diferentes. Desta forma, há uma relação anafórica, mas não uma relação de correferência. Mitkov (2002) classifica este tipo de anáfora como anáfora de **identidade-de-sentido** (*identity-of-sense*). Esse fenômeno ocorre quando a anáfora não se refere ao mesmo evento ou entidade que seu antecedente, mas a conceitos, eventos/entidades ou classe de eventos/entidades semelhantes (Mitkov, 2002), como é o caso do exemplo *n* - em que, apesar de não se tratarem do mesmo evento no modelo de discurso, o atraso de Márcia e o atraso de Márcio são eventos análogos, com características semelhantes.

Figura 7 – Ilustração da relação de anáfora sem correferência no exemplo *m*



Fonte: autora.

O mesmo acontece frequentemente em situações em que o antecedente de uma anáfora é um sintagma nominal marcado por um quantificador (todos, alguns, nenhum, a maioria, etc.) (Mitkov, 2002), como no exemplo *n*:

n) "[Toda mãe]$_i$ quer ver [seu]$_i$ filho feliz."

Em que "seu" é um pronome possessivo que retoma "toda mãe". Neste caso, um teste de substituição para confirmar uma possível correferência geraria:

o) "Toda mãe quer ver o filho de toda mãe feliz."

A sentença *n*, evidentemente, não tem o mesmo significado que a sentença *o*. Desta forma, "toda mãe" e "seu", no exemplo *o*, têm relação anafórica mas não correferem, apesar de se referirem a conceitos semelhantes.

Outro tipo de relação anafórica sem correferência é aquela que ocorre nas anáforas indiretas (também denominadas *bridging* em algumas literaturas[8]) (Mitkov, 2002), exemplificadas pelos casos *p* e *q*.

p) "Fomos ao clube de tênis ontem, mas as quadras estavam alagadas por conta da chuva."
q) "Um dos melhores times do Flamengo foi o de 1978-1983 - o meia era um jogador excepcional."

Nestes casos, a anáfora e seu antecedente não correferem e nem se referem a conceitos análogos, como acontece na anáfora de identidade de sentido. Ao invés disso, se referem a entidades ou conceitos que possuem relações como as de parte-todo ou de associação a um conjunto (Mitkov, 2002). Em *p*, "o clube de tênis" pode ser considerado antecedente de "as quadras", uma vez que, com o uso de conhecimento de mundo - neste caso, porque se sabe que clubes de tênis têm quadras de tênis -, é possível inferir que "as quadras" são parte do clube mencionado. Em *q*, "o meia" se refere ao meio-campista do time do Flamengo mencionado anteriormente, e não a qualquer outro meio-campista - ou seja, trata-se de uma relação de associação a um conjunto. Dessa forma, a relação anafórica não é explícita e direta, mas pode

[8] Como em Jurafsky *et al.* (2024) e Poesio *et al.*, (2016).

ser inferida com base em conhecimento prévio ou especializado do domínio temático em que se dá o discurso (Mitkov, 2002).

A distinção entre anáfora direta e indireta nem sempre é clara, uma vez que o nível de conhecimento necessário para estabelecer a relação entre os elementos envolvidos pode variar gradativamente conforme o seu tipo de relação (especialização, generalização, etc.), estando sujeita à interpretação do analista (Mitkov, 2002).

2.2 CLASSIFICAÇÃO DO PROBLEMA DE RESOLUÇÃO DE CORREFERÊNCIA

Alguns estudos, como o de Poesio *et al.* (2016), optam pelo uso do termo "resolução de anáfora" em um sentido amplo, incluindo a resolução tanto de correferência quanto de anáfora. No presente trabalho, contudo, será feita a distinção entre a resolução de anáfora e a resolução de correferência.

A **resolução de anáforas**, conforme discutido por Mitkov (2002) e Liu *et al.* (2023), refere-se ao processo de determinação dos antecedentes de uma anáfora (anáfora como *anaphor*, referindo-se ao elemento anafórico, e não ao fenômeno anáfora). A **resolução de correferência**, por outro lado, é uma tarefa mais ampla, que engloba a identificação de quaisquer menções em um texto que apontem para a mesma entidade ou evento no modelo de discurso, ou seja, que tenham relação de correferência - mesmo sem a presença de relações anafóricas, apesar de estes dois fenômenos coocorrerem frequentemente (Jurafsky *et al.*, 2024). A resolução de correferência facilita o agrupamento de informações relacionadas em um texto e contribui na execução de tarefas subsequentes que dependem, em algum nível, de seus resultados, como “vinculação de entidades (Kundu *et al.*, 2018), reconhecimento de entidades nomeadas (Dai *et al.*, 2019), resposta a perguntas (Bhattacharjee *et al.*, 2020), análise de sentimentos (Krishna *et al.*, 2017; Mao e Li, 2021) e *chatbots* (Zhu *et al.*, 2018)” (Liu *et al.*, 2023). Estes processos de resolução de anáforas e de correferência englobam diversas subtarefas, como extração de menções, detecção de anaforicidade/referência de menções, seleção de antecedentes e criação de correntes de correferência (Poesio *et al.*, 2016; Yang *et al.*, 2022).

2.3 CORPORA

O Aprendizado de Máquina é uma área da Inteligência Artificial voltada ao desenvolvimento de algoritmos que melhoram seu desempenho a partir de experiências com dados anteriores (Pustejovsky *et al.*, 2013). De acordo com Pustejovsky *et al.* (2013), existem três tipos principais de algoritmos de aprendizado de máquina: **aprendizado supervisionado**, que consiste em técnicas envolvendo o treinamento do modelo por meio de exemplos contendo entradas e suas respectivas saídas desejadas; **aprendizado não supervisionado**, que identifica padrões em dados não rotulados; e **aprendizado semi-supervisionado**, que combina ambos os métodos. Um **corpus** é "uma coleção de textos legíveis por máquina que foram produzidos em um contexto comunicativo natural" (Pustejovsky *et al.*, 2013, tradução nossa). Na área do PLN, o aprendizado supervisionado tem por elemento central corpora anotados com informações relevantes para a tarefa a ser aprendida (Pustejovsky *et al.*, 2013). Esses corpora anotados servem, também, para a avaliação e comparação do desempenho de sistemas de PLN (Poesio *et al.*, 2016).

De acordo com Poesio *et al.* (2016) (tradução nossa):

> Na década de 1990, o desejo de utilizar a resolução de anáfora[9] em aplicações práticas, especialmente no campo ainda emergente de extração de informações, levou a uma mudança de foco na pesquisa em resolução de anáfora para uma abordagem mais empírica do problema. Esse foco mais empírico também resultou na criação dos primeiros corpora anotados de tamanho médio, que possibilitaram o desenvolvimento baseado em dados de procedimentos de resolução e abordagens de aprendizado de máquina.

Conforme exposto em Poesio *et al.* (2016), o início do uso de corpora anotados no campo da resolução de correferência se deve, principalmente, às *Message Understanding Conferences* (MUC) (Kaufmann *et al.*, 1996 *apud* Poesio *et al.* 2016) - conferências que tinham por objetivo avaliar a qualidade de sistemas de extração de informação com base na comparação de seus resultados na resolução de corpora anotados. Em suas edições MUC-6 (1996) e MUC-7 (1998), foram introduzidas tarefas específicas de RC. Para tais tarefas, foram elaborados e disponibilizados aos participantes corpora anotados com informação de correferência, que possibilitaram e deram início a uma tendência para o desenvolvimento de abordagens mais

[9] Novamente, na obra de Poesio *et al.* (2016), o termo "resolução de anáfora" é utilizado em seu sentido amplo - incluindo além de todas as variedades de processamento de anáfora, a resolução de correferência e outras tarefas auxiliares (Poesio *et al.*, 2016, p. 2)

empíricas e baseadas em dados para a RC - em contraste com os métodos anteriores, que dependiam principalmente de regras manuais e heurísticas. Desde então, outros corpora anotados com informações de correferência vem sendo desenvolvidos, assumindo importância central para o avanço de métodos de aprendizado supervisionado nessa área e fornecendo uma base comum para o treinamento e avaliação de modelos de RC. (Poesio *et al.*, 2016).

Em revisão de literatura da área, Liu *et al.* (2023) destacam alguns dos corpora anotados com informações de correferência mais relevantes atualmente - em sua maioria do inglês. Dentre eles estão os corpora da tarefa compartilhada CoNLL 2012[10] (Pradhan *et al.*, 2012), desenvolvidos a partir do projeto OntoNotes[11] (Hovy *et al.*, 2006 *apud* Pradhan *et al.*, 2012) e amplamente utilizados para o treinamento e testagem de sistemas de resolução de correferência de entidades. A tarefa incluiu corpora em inglês, chinês e árabe, compostos por textos oriundos de diversas fontes (como notícias, conversas telefônicas e weblogs), com um total de 2802 documentos de treinamento (Pradhan *et al.*, 2012). Outro corpus destacado é o GAP (Webster *et al.*, 2018 *apud* Liu *et al.*, 2023), composto por trechos extraídos da Wikipédia e com enfoque em questões de ambiguidade de gênero, com 8908 pares de pronomes ambíguos e nomes, atuando como referência para testar vieses de gênero em sistemas de Inteligência Artificial (Liu *et al.*, 2023).

Além desses, há ainda outros exemplos destacados relevantes para a área com diferentes enfoques, como o ACE 2005, que contém aproximadamente 1800 arquivos de textos em inglês, chinês e árabe, com enfoque em entidades, relações e eventos; o LitBank (Bammam *et al.*, 2020, *apud* Liu *et al.*, 2023), corpus literário com 100 textos, anotados com categorias como pessoas, locais e organizações, utilizado para avaliação de correferências em contextos narrativos; o ECB+ (Cybulska *et al.*, 2014, *apud* Liu *et al.*, 2023), com 976 documentos, que se concentra na resolução de correferências de eventos entre múltiplos documentos e é utilizado para análises detalhadas de eventos em diferentes contextos; e o WEC (Wikipedia Event Coreference) (Eirew *et al.*, 2021, *apud* Liu *et al.*, 2023), que se destaca por reunir mais de 40 mil menções de eventos agrupados extraídos da Wikipédia, contribuindo para a resolução de correferências de eventos em larga escala (Liu *et al.*, 2023).

[10] Tarefa compartilhada envolvendo a resolução de correferência em inglês, chinês e árabe (Pradhan *et al.*, 2012).
[11] Projeto com o objetivo de anotar um corpus extenso, composto por diferentes gêneros textuais, integrando anotações de diferentes níveis da estrutura semântica de textos em inglês, chinês e árabe e contendo, entre outras informações, anotações de correferência (Hovy *et al.*, 2006 apud Pradhan *et al.*, 2012).

Estes corpora, assim como os demais corpora destacados por Liu *et al.* (2023), juntamente com seus respectivos idiomas, enfoques e tamanhos, podem ser conferidos na Tabela 1.

Tabela 1 – Exemplos de corpora relevantes do inglês e seus respectivos enfoques e tamanhos
(continua)

Corpus	Idioma	Enfoque	Tamanho (número de textos)			
			Treinamento	Validação	Teste	Total
CoNLL 2012 (Pradhan *et al*., 2012)	Inglês, chinês e árabe	Corpus de tarefa compartilhada	2802	343	348	3493
GAP (Webster *et al*., 2018)	Inglês	Viés de gênero em resolução de correferência pronominal	4000	908	4000	8908
KBP 2017 (Mitamura *et al*., 2017)	Inglês, chinês e espanhol	Correferência intra-documental de eventos			167	167
Automatic Content Extraction 2005	Inglês, chinês e árabe	Correferência intra-documental de eventos	529	28	40	599
LitBank (Bamman *et al*., 2020).	Inglês	Correferência intra-documental de longa distância				100
Winograd Schema Chalenge (Levesque, 2011)	Inglês	Conhecimento de senso comum em resolução de correferência pronominal	544	104	146	804
Definite Pronoun Resolution (Rahman *et al*., 2012)	Inglês	Casos complexos de pronomes definidos	1322		564	1886
Pronoun Disambiguation Problem (Davis *et al*., 2017)	Inglês	Conhecimento de senso comum em resolução de correferência pronominal				60
Winogender (Rudinger *et al*., 2018)	Inglês	Viés de gênero em resolução de correferência pronominal				720

Tabela 1 - Exemplos de corpora relevantes do inglês e seus respectivos enfoques e tamanhos (conclusão)

Corpus	Idioma	Enfoque	Tamanho (número de textos)			
			Treinamento	Validação	Teste	Total
WinoBias (Zhao *et al.*, 2018)	Inglês	Viés de gênero ocupacional em resolução de correferência pronominal	1580		1580	3160
KnowRef (Emami *et al.*, 2019)	Inglês	Casos desafiadores em resolução de correferência pronominal	7455		1269	8724
WikiCoref (Ghaddar *et al.*, 2016)	Inglês	Correferência na Wikipedia				30
Extension to Event Coreference Bank (Cybulska *et al.*, 2014)	Inglês	Correferência inter-documental	574	196	206	976
Richer Event Description (O'Gorman *et al.*, 2016)	Inglês	Correferência intra-documental de eventos				95
Georgetown University Multilayer Corpus (Zeldes, 2017)	Inglês	Corpus de tarefa compartilhada				168
Wikipedia Event Coreference (Eirew *et al.*, 2021)	Inglês	Correferência inter-documental de eventos	40529	1250	1893	43672
EmailCoref (Dakle *et al.*, 2020)	Inglês	Correferência de entidades em conversas por e-mail	36		10	46
BUG (Levy *et al.*, 2021)	Inglês	Viés de gênero em resolução de correferência pronominal			108K	108K

Fonte: Liu *et al.* (2023, tradução nossa).

2.3.1 Corpora do português

Enquanto que, atualmente, os maiores e mais relevantes corpora anotados com informação de correferência são predominantemente da língua inglesa (Liu *et al.*, 2023), na língua portuguesa os recursos para a tarefa de RC são consideravelmente mais escassos e menores (Fonseca *et al.*, 2018). Atualmente, os corpora do português anotados com algum tipo de informação de correferência são os corpora da primeira e segunda edições da campanha HAREM (Santos *et al.*, 2006; Freitas *et al.*, 2010), o corpus de Garcia *et al.* (2014), o Summ-it (Collovini *et al.*, 2007), o Summ-it++ (Antonitsch *et al.*, 2016) e o Corref-PT (Vieira *et al.*, 2018). Uma comparação entre o escopo e tamanho dos três maiores e mais relevantes corpora do português, o de Garcia *et al.* (2014), o Summ-it++ e o Corref-PT, pode ser observada na Tabela 2.

HAREM é uma campanha de avaliação de sistemas de reconhecimento de entidades nomeadas para o português desenvolvida pela iniciativa Linguateca[12], centro de recursos de PLN do português. A coleção dourada de textos anotados de sua primeira edição inclui 129 textos de diferentes gêneros (como jornal, escritos técnicos, e-mails, textos políticos e discursos orais), totalizando 5.132 entidades anotadas (Santos *et al.*, 2006). Em sua segunda edição, foram adicionados a essa coleção dourada textos de novos gêneros textuais (como textos de *blogs* e da Wikipédia), ampliando o total de entidades anotadas para 7.847. O corpus conta com textos de Portugal, Brasil, Moçambique e Angola e contém, além de informações manualmente anotadas das entidades (incluindo sua distribuição em diferentes categorias semânticas), suas relações de identidade (Freitas *et al.*, 2010).

O corpus criado por Garcia *et al.* (2014), por sua vez, trata-se de um conjunto de corpora multilíngues com anotações de correferência de entidades pessoais no português, galego e espanhol. O corpus é manualmente anotado seguindo o formato da tarefa compartilhada #1 do SemEval-2010[13] (Recasens *et al.*, 2010 *apud* Garcia *et al.*, 2014). Os corpora contêm textos jornalísticos e da Wikipédia. Cada um deles, português, galego e espanhol, contém entre 42 e 51 mil *tokens*, com os três corpora contabilizando um total de 11.995 menções (Garcia *et al.*, 2014).

---

[12] Disponível em: https://www.linguateca.pt/

[13] RECASENS, Marta *et al.* Semeval-2010 task 1: Coreference resolution in multiple languages. In: **Proceedings of the 5th international workshop on semantic evaluation**. 2010. p. 1-8.

O corpus Summ-it++ é uma versão aprimorada do corpus Summ-it, desenvolvido, prioritariamente, para o estudo da tarefa de sumarização automática. O Summ-it é composto por cinquenta textos jornalísticos do caderno de ciências do jornal Folha de São Paulo, contendo, cada um, entre 127 e 654 palavras, e anotados em diferentes camadas de informação (Collovini *et al.*, 2007). Os textos foram anotados com informações de correferência entre sintagmas nominais de forma manual, seguindo as diretrizes elaboradas por Coelho *et al.* (2006, *apud* Antonitsch *et al.*, 2016), e contêm um total de aproximadamente 20 mil *tokens* e 560 cadeias de correferência. O Summ-it++ adiciona ao Summ-it duas novas camadas semânticas, de entidades nomeadas e de suas relações semânticas, e converte o corpus para o formato SemEval (Recasens *et al.*, 2010, *apud* Antonitsch *et al.*, 2016). (Antonitsch *et al.*, 2016)

O Corref-PT, por sua vez, é um corpus do português semiautomaticamente anotado com informações de correferência entre sintagmas nominais. Inclui um total de 182 textos, em sua maioria notícias, mas também artigos da Wikipédia, contabilizando um total de 3898 correntes de correferência. É disponibilizado no formato Semeval e XML. As anotações foram realizadas em duas fases: uma fase automática inicial utilizando a ferramenta de anotação automática de correferência CORP (Fonseca *et al.*, 2017), seguida por uma revisão manual por meio da ferramenta de edição CorrefVisual[14] (Tubino *et al.*, 2015, *apud* Vieira *et al.*, 2018) - desenvolvida para possibilitar a edição das correntes anotadas pelo CORP (Vieira *et al.*, 2018).

Tabela 2 – Comparação entre os 3 mais relevantes corpus do português anotados com informações de correferência

Corpus	Escopo	Número aproximado de *tokens*	Textos
Garcia *et al.* (2014)	Correferência entre entidades pessoais	51K	97
Summ-it++	Correferência entre sintagmas nominais	20K	50
Corref-PT	Correferência entre sintagmas nominais	124K	182

Fonte: autora.

[14] Disponível para acesso em: https://www.inf.pucrs.br/linatural/wordpress/recursos-e-ferramentas/correfvisual/

2.4 MODELOS DE RESOLUÇÃO DE CORREFERÊNCIA

Diferentes abordagens podem fundamentar os sistemas de RC. Conforme exposto em Poesio *et al.* (2016), o **modelo de pares de menções** (*mention-pair model*) é uma das abordagens mais clássicas à tarefa, que segue sendo amplamente utilizada. Nessa abordagem, a RC é organizada como um problema de classificação binária: o modelo analisa pares de menções, determinando se cada par se refere ou não à mesma entidade. Embora eficaz em sua simplicidade, o modelo de pares de menções apresenta algumas limitações. Sua atuação se restringe a um processo de tomada de decisões de caráter local, em que pares de menções são avaliados isoladamente, resultando em uma capacidade limitada de considerar o contexto discursivo mais amplo do material analisado. Além disso, como cada candidato a antecedente de uma menção é considerado de maneira independente dos demais, o modelo avalia apenas a qualidade de um antecedente em relação à menção ativa, sem compará-lo aos demais candidatos e às demais decisões já feitas sobre outros pares, o que restringe sua capacidade de expressividade. (Poesio *et al.*, 2016)

Para superar essas limitações, foram propostos modelos que expandem suas capacidades. O **modelo entidade-menção** (*entity-mention model*) considera em suas análises grupos de menções, e não apenas pares isolados. Assim como o modelo de pares de menções, esse modelo também aborda a RC como uma tarefa de classificação binária. Contudo, esse modelo também constrói, de maneira dinâmica, cadeias de correferência - grupos contendo cada menção identificada a uma dada entidade -, incorporando ao processo de tomada de decisão as informações sobre menções previamente processadas e agrupadas, em um processo incremental. Isso permite que o modelo assuma uma visão mais global do texto quando comparada à do modelo de pares de menções. (Poesio *et al.*, 2016)

Apesar do aprimoramento da perspectiva global em relação ao modelo de pares de menções, o modelo entidade-menção não emprega um método que permita a ponderação simultânea de todos os candidatos na escolha do candidato a antecedente mais provável de uma menção - apenas uma decisão embasada no conjunto de todas as decisões anteriores. Diante dessa questão, foi proposto o **modelo de ranqueamento de menções** (*mention-ranking model*). Esse modelo permite que todos os potenciais antecedentes de uma menção sejam considerados simultaneamente. Em vez de tratar cada par de forma independente, o modelo ranqueia todos os candidatos possíveis de cada menção e seleciona o mais provável dentre eles. (Poesio *et al.*, 2016)

O **modelo de ranqueamento de *clusters*** (*cluster-ranking model*), por sua vez, é uma abordagem mais avançada para a tarefa de RC, proposta por Rahman *et al.* (2009, *apud* Poesio *et al.*, 2016). Embora o modelo de ranqueamento de menções permita que todos os candidatos a antecedente de uma dada menção sejam ranqueados e comparados simultaneamente, ele não possibilita o uso das informações a nível de correntes de correferência, que permitem decisões embasadas no conjunto de decisões anteriores; por outro lado, embora o modelo de menção-entidade permita o uso dessas informações, ele não permite a ponderação simultânea de todos os candidatos a antecedente. Dessa forma, ao invés de avaliar pares de menções, o modelo de ranqueamento de *clusters* recorre a um sistema de ranqueamento para decidir a melhor corrente de correferência, e não menção isolada, a qual uma dada menção deve ser vinculada. Essa abordagem combina os pontos fortes do modelo de ranqueamento de menções, que permite ordenar os candidatos a antecedentes, e do modelo entidade-menção, que acessa as informações de correntes de correferência. (Poesio *et al.*, 2016)

De acordo com Liu *et al.* (2023), nos últimos dez anos, o foco da pesquisa evoluiu de modelos que utilizam características de entidades e aprendizado de máquina tradicional para modelos de aprendizado profundo que se valem de informações contextuais e explícitas de bases de conhecimento. Posteriormente, a atenção se voltou para métodos desenvolvidos sobre grandes modelos pré-treinados com base na arquitetura de *transformers*. A seguir, será realizada uma breve recapitulação do desenvolvimento da resolução de correferência na última década, fundamentada na revisão de escopo conduzida por Liu *et al.* (2023).

2.4.1 Modelos baseados em características

Os modelos baseados em características utilizam informações linguísticas como *tags* de parte do discurso, reconhecimento de entidades nomeadas e informações semânticas de nível superficial. Predominantes nas fases iniciais da pesquisa em RC, esses modelos dependem de técnicas de aprendizado de máquina tradicional, em vez de métodos de aprendizado profundo (Liu *et al.*, 2023). De acordo com Liu *et al.* (2023, tradução nossa):

> As características mais comumente utilizadas nesse tipo de abordagem incluem palavras de opinião (Ding e Liu 2010), tags de parte do discurso (Ding e Liu 2010; Atkinson *et al.* 2015), segmentação de texto (Ding e Liu 2010), tags de reconhecimento de entidades nomeadas (Atkinson *et al.* 2015; Raghunathan *et al.* 2010a), informações semânticas (Durrett e Klein 2014), funções sintáticas (Durrett e Klein 2014), posições das palavras e palavras-chave (Durrett e Klein 2014; Raghunathan *et al.* 2010a).

2.4.2 Modelos baseados em redes neurais recorrentes

Com o passar do tempo, os modelos baseados em características foram superados por modelos baseados em aprendizado profundo de máquina. Boa parte dos modelos dessa geração baseiam-se no uso de Redes Neurais Recorrentes (RNRs). As RNRs são um tipo de rede neural projetada para trabalhar com dados sequenciais, utilizando uma estrutura de loops internos para manter uma memória dos estados anteriores, e considerando tais estados para a codificação de cada nova palavra. Sendo assim, as RNRs processam as entradas de maneira sequencial, uma de cada vez, e dependem do processamento de todos os estados anteriores para o processamento do próximo - diferentemente dos métodos de aprendizado de máquina tradicional, como regressão linear ou redes neurais clássicas, que tratam cada entrada de forma independente, sem considerar dependências temporais ou sequenciais entre os dados. Os modelos baseados em RNRs treinam essas redes para considerar informação contextual, e costumam não incluir conhecimento externo (como bases de dados semânticos ou outras ferramentas semelhantes) além do próprio conjunto de dados de treinamento. (Liu *et al.*, 2023)

Alguns modelos, como o de Clark e Manning (2015 *apud* Liu *et al.*, 2023), adotam uma abordagem baseada em entidades, ponderando pontuações de pares de menções para atribuir propriedades às entidades referidas pelas menções de cada corrente de correferência, que, por sua vez, influenciam as pontuações dos próximos pares analisados. Outro exemplo é o modelo de Lee *et al.* (2017 *apud* Liu *et al.* 2023), que construíram um modelo de RC *end-to-end* - ou seja, que realiza a tarefa integralmente, dispensando o uso de ferramentas como um *parser* sintático ou um detector de menções explícito. Esse modelo considera todos os *spans* em um texto como possíveis menções e aprende a distribuição de probabilidade sobre os possíveis antecedentes dos mesmos, usando representações de limites baseadas em contexto e um mecanismo de atenção para encontrar o núcleo desses *spans*. Posteriormente, Lee *et al.* (2018 *apud* Liu *et al.* 2023) melhoraram esse modelo utilizando um mecanismo de atenção para atualizar representações de *spans* e permitir que decisões futuras de correferência se baseiem no conjunto das decisões anteriores. (Liu *et al.*, 2023)

2.4.3 Modelos baseados em conhecimento

Os modelos baseados em conhecimento são semelhantes aos modelos contextuais baseados em redes neurais, mas além dos conjuntos de dados de treinamento, esses modelos

empregam, explicitamente, conhecimento externo, como bases de conhecimento de senso comum ou especializado. Esses modelos treinam RNRs para incorporar e utilizar essas informações adicionais. (Liu *et al.*, 2023)

Diferentes estudos têm explorado características variadas para esse tipo de abordagem. Aralikatte *et al.* (2019 *apud* Liu *et al.* 2023) integraram informações de conhecimento da Wikipedia e Wikidata em modelos de aprendizado por reforço que utilizam a comparação das relações obtidas por meio da resolução de correferência a uma base de conhecimento como indicador indireto de seu desempenho nessa tarefa. Emami *et al.* (2018 *apud* Liu *et al.* 2023) desenvolveram um sistema automatizado que utiliza um módulo de busca de conhecimento para resolver tarefas de RC complexas, gerando consultas para mecanismos de busca com base nas necessidades impostas por cada *input*, e categorizando as informações retornadas para resolver os problemas de correferência. Zhang *et al.* (2019 *apud* Liu *et al.* 2023) exploraram o uso de conhecimento externo e optaram pela incorporação de informação em *triplets*[15] no lugar da incorporação por regras - uma abordagem mais atualizada que facilita a generalização do modelo. (Liu *et al.*, 2023)

2.4.4 Modelos baseados em *transformers*

Transformers são uma arquitetura de redes neurais proposta por Vaswani *et al.* (2017 *apud* Liu *et al.* 2023) que trouxe uma mudança paradigmática para o campo do PLN e outras áreas da inteligência artificial. Fundamentados em mecanismos de autoatenção, esses modelos permitem a análise simultânea de todas as palavras de uma sequência de dados, como frases ou parágrafos. Em contraste com as RNRs, que executam operações de codificação de maneira sequencial e iterativa, os *transformers* possibilitam a paralelização dessas operações, resultando em um aumento significativo em sua velocidade. Esta abordagem permite uma percepção mais global dos dados processados e a identificação de relações de longo alcance que RNRs, devido à sua natureza mais local e restrita, frequentemente não conseguem captar. Modelos de Linguagem de Grande Escala (LLMs), como o BERT e os GPTs (OpenAI 2022; 2023 *apud* Liu *et al.* 2023), são exemplos de *transformers* que foram treinados em vastos conjuntos de dados

[15] Método de representação de uma informação textual em menos palavras sem perder o contexto. São normalmente compostos por um sujeito, um predicado e um objeto, formando uma estrutura que expressa uma informação ou fato.

textuais para realizar uma variedade de tarefas, incluindo a geração de respostas coerentes, compreensão de linguagem e tradução automática. (Liu *et al.*, 2023).

Com a introdução da arquitetura de *transformers*, houve um surgimento de um número significativo de modelos baseados em BERT para a resolução de correferência. Joshi *et al.* (2019 *apud* Liu *et al.* 2023), por exemplo, substituíram completamente o codificador baseado em LSTM por BERT no modelo de Lee *et al.* (2018 *apud* Liu *et al.* 2023), obtendo resultados consideravelmente melhores. Joshi *et al.* (2020 *apud* Liu *et al.* 2023) propuseram o *span*BERT, que possui a mesma arquitetura que o BERT, mas com treinamento específico para a tarefa de RC, e o utilizaram no lugar do BERT como codificador no modelo de seu trabalho anterior (Joshi *et al.*, 2019 *apud* Liu *et al.* 2023). Wu *et al.* (2020 *apud* Liu *et al.* 2023) apresentaram o CorefQA, modelo em que a resolução de correferência é definida como um problema de previsão de menções em um contexto de resposta a perguntas, e que permite a recuperação de menções que foram perdidas durante a fase de reconhecimento de menções. (Liu *et al.*, 2023)

2.5 TRABALHOS RELACIONADOS

2.5.1 Modelos de resolução de correferência em português

No português, existem alguns modelos de resolução de correferência elaborados com e sem aprendizado supervisionado e, em sua maioria, baseados em regras e/ou conhecimento semântico. O modelo proposto por Fonseca *et al.* (2016) utiliza aprendizado supervisionado e incorpora conhecimento semântico às características extraídas das menções, como relações de sinonímia, hiponímia e hiperonímia, além de categorias de entidades (pessoa, lugar, organização), obtidas de recursos como a base de dados semânticos Onto-PT (Oliveira *et al.*, 2014 *apud* Fonseca *et al.*, 2016) e o as anotações semânticas obtidas através do parser Palavras (Bick *et al.*, 2000 *apud* Fonseca *et al.*, 2016). Utiliza a abordagem de pares de menções para a resolução das correferências, e a classificação dos pares é feita com base em um conjunto de características que incluem tanto aspectos básicos, como a correspondência exata de cadeias de caracteres e a distância entre os NPs, quanto características semânticas mais complexas, como a relação hierárquica entre termos e a similaridade semântica (Fonseca *et al.*, 2016).

Os experimentos realizados pelos autores avaliaram seis variações de modelos, cada um incorporando diferentes características semânticas elencadas. O primeiro modelo, chamado de *Baseline*, utiliza apenas as características não semânticas, servindo como referência inicial.

O segundo modelo, denominado *EntityCat*, acrescenta ao *Baseline* as características relacionadas às categorias de entidade. O terceiro modelo adiciona a característica de sinonímia ao *EntityCat*. Em seguida, o quarto modelo incorpora a relação de hiponímia, enquanto o quinto adiciona a relação de hiperonímia à base do *EntityCat*. Por fim, o último modelo (*Full semantic*) combina todas as cinco características semânticas (categorias de entidade, sinonímia, hiponímia e hiperônimo) para avaliar o impacto combinado dessas informações no desempenho do sistema de resolução de correferência. Os modelos foram testados no corpus Summ-it e avaliados utilizando os índices de precisão, *recall* e F^{1}[16] para as respostas positivas e negativas, respectivamente. (Fonseca *et al.*, 2016)

Os resultados obtidos podem ser conferidos na Tabela 3.

Tabela 3 – Resultados dos modelos propostos por Fonseca *et al.* (2016)

Modelo	Média Prec. Pos.	Média Prec. Neg.	Média *recall* Pos.	Média *recall* Neg.	Média F^{1} Pos.	Média F^{1} Neg.
Baseline	79,16%	64,99%	53,81%	**85,66%**	63,96%	73,87%
EntityCat	78,57%	66,37%	**57,26%**	84,24%	66,16%	74,21%
EntityCat + Sinonímia	78,61%	66,36%	57,20%	84,30%	66,13%	74,22%
EntityCat + Hiponímia	79,73%	**66,59%**	57,13%	85,41%	**66,52%**	74,82%
EntityCat + Hiperonímia	79,04%	66,36%	56,99%	84,73%	66,13%	74,39%
Full semantic	**79,92%**	66,56%	56,95%	85,62%	66,45%	**74,88%**
Coeficiente de variação	2,68%	1,17%	4,65%	3,12%	1,85%	**1,06%**
Desvio padrão	2,13	**0,77**	2,63	2,65	1,22	0,79

Fonte: Fonseca *et al.* (2016) - tradução nossa.

Posteriormente, Fonseca *et al.* (2017) exploram diferentes combinações de regras lexicais, sintáticas e semânticas, desenvolvendo a ferramenta de RC em português CORP. O modelo segue uma arquitetura multipassos, em que cada etapa consiste na aplicação de uma das regras desenvolvidas para agrupar ou não um par de menções. As regras incluem análises de casamento de padrões exato, casamento parcial pelo núcleo, acrônimo, predicado nominativo, pronome relativo, casamento restrito pelo núcleo, modificadores compatíveis, encapsulamento

[16] Ver seção 2.6, "Métricas e *scorers*".

de menções, casamento entre nomes próprios, casamento parcial entre nomes próprios e regras semânticas de hiponímia e sinonímia. Utilizam o recurso CoGrOO (Silva, 2013 *apud* Fonseca *et al.* 2017) para anotação de sintagmas nominais, análise morfológica, lematização e detecção de menções. Foi utilizada, também, a ontologia Onto.PT para extrair as relações semânticas entre pares de menções (Fonseca *et al.* 2017).

O modelo foi avaliado na resolução do corpus Summ-it++ por meio das métricas MUC, B^3 e CoNLL[17], e os resultados obtidos podem ser conferidos na Tabela 4.

Tabela 4 – Resultados obtidos pelo modelo proposto por Fonseca *et al.* (2017)

Modelo	MUC			B^3			CEAFe			CoNLL
	Prec.	Rec.	F^1	Prec.	Rec.	F^1	Prec.	Rec.	F^1	F^1
Fonseca *et al.* (2017)	42,3	53,6	47,3	38,7	50,8	43,9	45,6	52,8	48,9	46,7

Fonte: Fonseca *et al.* (2017).

O sistema proposto por Rocha *et al.* (2017), por sua vez, é baseado em aprendizado de máquina e segue a abordagem de pares de menções, em que o problema é tratado como uma tarefa de classificação binária. Nesse sistema, a RC é feita por modelos treinados com base em diferentes conjuntos de treinamento consistindo em pares de menções gerados a partir do corpus Summ-it++, classificados como positivos (correferentes) ou negativos (não correferentes). Os métodos heurísticos utilizados para a criação dos conjuntos de treinamento foram os de todos os antecedentes, antecedente mais próximo, antecedente mais confiável e vizinhos antecedentes mais confiáveis (MCAN).

Para o método de **todos os antecedentes**, um conjunto de pares positivos é criado com quaisquer menções correferentes m_i e *me*, e um conjunto de pares negativos é criado com quaisquer menções m_i e m_e não correferentes. Esse método gera um conjunto de treinamento desbalanceado, uma vez que o número de pares não correferentes é naturalmente maior que o de pares correferentes. Os demais métodos explorados visam mitigar esse desbalanceamento. No método de **antecedente mais próximo**, para cada menção m_e, um par anotado como correferente é formado com seu antecedente mais próximo, m_i, e pares anotados como não correferentes são criados com m_e e cada menção entre m_i e m_e. No método de **antecedente mais confiável**, para cada menção m_e, caso esta seja uma menção não-pronominal, um exemplo positivo é gerado entre m_e e seu antecedente não-pronominal mais próximo, m_i; e, para os

[17] Ver seção 2.6, "Métricas e *scorers*".

demais casos$_e$, um exemplo positivo é formado entre m_e e seu antecedente mais próximo (pronominal ou não) m_i. Os exemplos negativos são gerados entre a menção m_e e cada menção entre m_i e m_e (assim como no método de antecedente mais próximo). O método **MCAN** gera os pares negativos de forma semelhante ao método de antecedente mais próximo, com a diferença de que, ao invés de gerar um par negativo entre m_e e cada menção ocorrendo entre m_e e m_i (seu antecedente mais próximo), gera apenas um número pré-definido de pares negativos entre essas menções. Para gerar os pares positivos, segue as restrições utilizadas pelo método de antecedente mais confiável. Além disso, foram também utilizados métodos de balanceamento randômicos (***Random Undersampling 1*** e ***Random Undersampling 2***), que removem, de forma aleatória, alguns dos pares negativos do conjunto de treinamento formado pelo método de todos os antecedentes.

Os modelos treinados decidem se um cada par de menções é correferente ou não utilizando um conjunto de características lexicais, sintáticas, morfológicas e semânticas. O texto é processado da esquerda para a direita, e para cada menção ativa, pares de menções são gerados da direita para a esquerda, com todas as menções anteriores, e são então apresentados ao classificador de correferência. O classificador decide se cada par de menções apresentado é correferente ou não, e o processo para assim que um antecedente é encontrado ou o início do texto é alcançado. No final, menções conectadas formam cadeias de correferência, enquanto menções sem conexões formam *singletons*. Os modelos foram testados com e sem o uso de características baseadas em semântica e foram, então, avaliados com base nos índices de precisão, *recall* e F^1 para a classificação de pares de menções correferentes e não correferentes, respectivamente, de um conjunto de teste composto por pares extraídos do corpus *Summ-it++*. (Rocha *et al.*, 2017).

Na Tabela 5, é possível observar os resultados obtidos para os modelos treinados com os conjuntos de cada método de criação de conjuntos de treinamento, em suas variações com e sem o uso de características semânticas.

Tabela 5 – Resultados obtidos pelo modelo proposto por Rocha *et al.* (2017)

(continua)

Método	Pares correferentes				Pares não correferentes			
	Número de instâncias	Prec.	Rec.	F^1	Número de instâncias	Prec.	Rec.	F^1
Sem características semânticas								
Todos os antecedentes	3320	0,72	0,20	0,31	38759	0,94	0,99	0,96

Tabela 5 – Resultados obtidos pelo modelo proposto por Rocha *et al.* (2017)
(conclusão)

Método	Pares correferentes				Pares não correferentes			
	Número de instâncias	Prec.	Rec.	F¹	Número de instâncias	Prec.	Rec.	F¹
Antecedente mais confiável	1267	0,84	0,27	0,40	8812	0,90	0,99	0,95
Antecedente mais próximo	1273	0,86	0,25	0,39	8701	0,90	0,99	0,95
MCAN	2871	0,64	0,41	0,50	4351	0,69	0,85	0,76
Com características semânticas								
Todos os antecedentes	3320	0,76	0,23	0,35	38759	0,94	0,99	0,96
Antecedente mais confiável	1267	0,78	0,41	0,54	8812	0,92	0,98	0,95
Antecedente mais próximo	1273	0,75	0,40	0,52	8701	0,92	0,98	0,95
MCAN	2871	0,79	0,51	0,62	4351	0,74	0,91	0,82
Random Undersample 1	1267	0,74	0,31	0,44	8812	0,91	0,98	0,94
Random Undersample 2	2871	0,74	0,53	0,62	4351	0,74	0,88	0,80

Fonte: Rocha *et al.* (2017).

Mais tarde, Fonseca *et. al* (2018) propuseram um modelo de resolução de correferências nominais no português baseado na incorporação de regras sintático-semânticas à ferramenta CORP. O sistema começa a tarefa pela extração de sintagmas nominais e seus atributos por meio do uso do parser CoGrOO, seguido por um pré-processamento que elimina sintagmas nominais que começam com entidades numéricas - uma vez que estes não são abordados pelo modelo. Em seguida, é aplicado o conjunto de regras sintático-semânticas desenvolvido. Relações semânticas, como as de sinonímia e hiponímia, são identificadas utilizando a ferramenta Onto.PT, e as entidades são categorizadas com o auxílio da ferramenta Repentino (Sarmento *et al.*, 2006, *apud* Fonseca *et al.*, 2016) e listas pré-elaboradas de tipos de entidade. As regras são aplicadas para classificar os pares como correferentes ou não, e o modelo utiliza uma estrutura de grafos para armazenar as informações desse processamento. Em seguida, é aplicado um método de agrupamento que determina se uma menção se refere a

uma entidade já referida ou nova no discurso, elaborando assim as correntes de correferência. O sistema lida principalmente com correferências nominais, excluindo correferências numéricas devido à sua baixa frequência e necessidade de processamento específico. Ele trata de substantivos comuns e próprios, resolvendo correferências tanto para categorias gerais quanto para entidades específicas. O modelo foi avaliado nos corpus Summ-it++, Corref-PT e o de Garcia *et al.* (2014), utilizando as métricas MUC, B^3 e CoNLL[18]. (Fonseca *et. al*, 2018)

Os resultados obtidos podem ser conferidos na Tabela 6.

Tabela 6 – Resultados obtidos pelo modelo proposto por Fonseca *et al.* (2018)

Modelo	Corpus	MUC			B^3			CEAFe			CoNLL
		P	R	F^1	P	R	F^1	P	R	F^1	F^1
Sem informação semântica	Summ-it++	58.8	44.4	50.6	59.3	74.0	49.0	53.7	54.2	54.0	51.2
	Corref-PT	61.2	47.8	54.8	61.2	40.5	48.7	50.2	51.0	50.6	51.4
Com informação semântica	Summ-it++	45.1	52.1	48.3	43.8	49.0	46.5	45.7	57.4	50.9	48.6
	Corref-PT	54.9	50.2	52.5	51.8	43.6	47.3	46.2	52.8	49.3	49.7

Fonte: Fonseca *et al.* (2018).

Lima *et al.* (2018) investigaram o impacto da integração de duas bases semânticas diferentes à ferramenta CORP - a base Onto.PT, já utilizada com a ferramenta em Fonseca *et al.* (2017) e Fonseca *et al.* (2018), e ConceptNet, base de dados que oferece uma rede semântica global com relações entre palavras e conceitos. Foi implementado um mecanismo que possibilita a integração de novas bases semânticas ao CORP, permitindo a escolha das bases durante a execução e a unificação das relações entre diferentes bases. (Lima *et al.*, 2018)

Dois métodos de agrupamento de menções em correntes foram utilizados para a associação de menções às correntes de correferência. O método *Clustering A* conecta uma menção ao antecedente mais próximo quando ao menos uma das regras semânticas estabelecidas é satisfeita. O método *Clustering B* explora a representação do discurso, assumindo que uma menção é nova no discurso se não estiver ligada a nenhum antecedente disponível. Nesse caso, uma nova cadeia de correferência é criada. Quando mais de uma cadeia

[18] Ver seção 2.6, "Métricas e *scorers*".

existente pode ser aplicada, um critério de agrupamento é adotado para decidir a qual cadeia a menção será vinculada. (Lima *et al.*, 2018)

Os experimentos foram realizados no corpus Corref-PT, e os resultados foram avaliados utilizando as métricas MUC, B, CEAFe e CoNLL[19]. Foram testadas as variantes combinando o método de agrupamento de menções A e B com as bases semânticas Onto.PT e ConceptNet, tanto individualmente quanto de forma combinada (Lima *et al.*, 2018). Os resultados obtidos podem ser verificados na Tabela 7.

Tabela 7 – Resultados obtidos por Lima *et al.* (2018)

Método de agrupamento de menções	Base	MUC			B^3	CEAFe	CoNLL
		Prec.	Rec.	F^1	F^1	F^1	
*Cluster*ing A	Onto.PT	43,96%	51,9%	47,6%	40,05%	44,93%	44,19%
	ConceptNet	54,15%	49,58%	51,77%	45,27%	47,88%	48,31%
	Onto.PT + ConceptNet	43,85%	51,86%	47,52%	39,94%	44,88%	44,11%
*Cluster*ing B	Onto.PT	54,6%	49,93%	52,16%	47,19%	49,29%	49,55%
	ConceptNet	62,89%	47,83%	54,33%	48,54%	50,57%	51,15%
	Onto.PT + ConceptNet	54,52%	49,96%	52,14%	47,18%	49,96%	49,19%

Fonte: Lima *et al.* (2018).

Cruz *et al*. (2018) elaboraram diferentes modelos de RC baseados em aprendizado de máquina, utilizando redes neurais profundas, com abordagem de pares de menções. O sistema proposto explora o aprendizado por transferência entre o espanhol e o português, aproveitando os corpora AnCora (espanhol) e Corref-PT (português), além de vetores de palavras multilíngues. Esses vetores alinham os espaços semânticos de ambas as línguas, facilitando a transferência de conhecimento entre elas. O processo envolve treinar modelos de redes neurais com conjuntos de treinamento criados a partir dos corpora anotados - compostos por pares de cada uma das menções com todos seus candidatos a antecedentes e submetidos a diferentes métodos de balanceamento randômico. São exploradas técnicas de sub-amostragem para equilibrar esse conjunto de dados, que é naturalmente desbalanceado devido à predominância

[19] Ver seção 2.6, "Métricas e *scorers*".

de exemplos negativos em relação aos positivos. Três modelos são criados como referência para atestar a performance do modelo frente a um método de escolhas aleatórias: No método Rand1, cada menção tem 50% de chance de ser associada a um antecedente aleatório, com cada uma das menções anteriores tendo a mesma chance de ser selecionada como antecedente. Dessa forma, 50% das menções serão associadas a um antecedente, e as demais serão consideradas singletons. O modelo Rand2 segue a mesma lógica, mas, em vez de uma probabilidade fixa de 50%, a chance de que uma menção seja atrelada a um antecedente segue a proporção de menções correferentes no corpus. Assim, por exemplo, se 30% das menções no corpus forem correferentes, cada menção terá uma probabilidade de 30% de ser associada a um antecedente aleatório. Assim como no Rand1, a seleção do antecedente é uniformemente distribuída entre as menções anteriores. Também foi criado um método de referência heurístico, AlwaysNo, em que todas as menções são consideradas singletons e, portanto, todo par de menção é considerado não correferente.

Os modelos utilizam diferentes arquiteturas (representadas por *Arch1*, *Arch2*, *Arch2-dense*, *Arch-deep-CNN*, *Arch-biLSTM* na Tabela 8)[20] e usam um algoritmo de ligação determinístico que liga cada menção ao seu antecedente positivamente identificado mais próximo, caso haja um. As entradas dos modelos incluem representações vetoriais das menções e a distância entre elas em termos de tokens e sentenças, e cada modelo trabalha em duas etapas - sendo a primeira a extração das características representativas das menções, e a segunda a definição da afinidade de correferência entre elas. O desempenho dos modelos foi avaliado utilizando as métricas MUC, B^3, CEAFe, BLANC e CoNLL[21] e testado nos corpora AnCora e Corref-PT. (Cruz *et al.*, 2018). O resultado obtido por cada uma das arquiteturas testadas pode ser verificado na Tabela 8.

Tabela 8 – Resultados obtidos por Cruz *et al.* (2018) na resolução do corpus Corref-PT

(continua)

Modelo		MUC			B^3			CEAFe			BLANC			CoNLL
		Prec.	Rec.	F^{11}	Prec.	Rec.	F^{11}	Prec.	Rec.	F^{11}	Prec	Rec.	F^{11}	
PT	Rand1	13,7	20,5	16,5	30,5	54,2	39	46,2	24,1	31,7	50,4	5,7	50,4	29,1
	Rand2	3,6	11,7	5,4	27,2	79,5	40,6	51,7	17,6	26,2	50,1	51,4	49,2	24,1
	AlwaysNo	0	0	0	26,4	100	41,7	52,2	13,8	21,8	50	47,3	48,6	21,2
	Arch1	43,8	55,4	48,9	46	57,6	51,2	49,5	31,2	38,3	57,6	55,7	56,4	46,1
	Arch2	46,8	59,7	52,5	46,97	62,6	53,7	55,1	34,5	42,4	58,3	60,9	59,3	49,5
	Arch2-dense	46,7	59,2	52,1	48,1	59,3	52,9	51,2	22,6	39,7	60,1	58,6	59,1	48,2

[20] Para mais detalhes sobre as arquiteturas, consultar Cruz *et al.* (2018).
[21] Ver seção 2.6, "Métricas e *scorers*".

Tabela 8 – Resultados obtidos por Cruz *et al.* (2018) na resolução do corpus Corref-PT
(conclusão)

Modelo		MUC			B^3			CEAFe			BLANC			CoNLL
		Prec.	Rec.	$F1^1$	Prec.	Rec.	$F1^1$	Prec.	Rec.	$F1^1$	Prec	Rec.	$F1^1$	
PT	Arch-deep-CNN	41,8	53	46,7	44,7	58	50,4	50,8	32	39,2	64,6	65,1	57,1	45,5
	Arch-biLSTM	46,8	58,2	51,8	48,4	58,9	53,1	51,4	33,4	40,5	59,1	58,5	58,7	48,4
ES->PT	Arch1	0,6	46,7	1,2	26,7	99,5	42,1	52,2	14	22,1	50,1	66,4	48,7	21,8
	Arch2	56,9	60,9	58,7	58,6	39,7	45,8	33	28	29,7	52,3	50,6	41,2	44,8
	Arch2-dense	27,2	46,8	33,2	39,7	68,8	49,5	48,1	21,2	29,2	53,1	53,2	51,7	37,3
	Arch-deep-CNN	0,2	33,1	0,3	26,5	99,7	41,8	52,2	13,8	21,9	50	59,6	48,7	21,4
	Arch-biLSTM	5,4	40	9,6	28,1	92	43,1	51,9	15,2	23,5	50,8	56,2	50,5	25,4

Fonte: Cruz *et al.* (2018).

2.5.2 Uso de engenharia de *prompt* para a resolução de correferência

Atualmente, os modelos estado da arte em resolução de correferência são amplamente baseados em LLMs pré-treinados (Liu *et al.*, 2023). No entanto, a maioria desses modelos utiliza aprendizado supervisionado, que, apesar de proporcionar resultados bastante vantajosos, é altamente onerosa em termos de recursos (Yang *et al.*, 2022).

De acordo com Knoth *et al.* (2024), a engenharia de *prompt*, no contexto de interações com sistemas de IA baseados em LLMs, consiste na habilidade de formulação de instruções precisas e bem estruturadas, com o intuito de obter respostas ou informações desejadas, otimizando a eficácia da interação com o modelo de linguagem. De acordo com Yang *et al.* (2022), observa-se no campo do PLN uma tendência crescente para a utilização de engenharia de *prompt* em LLMs como alternativa ao aprendizado supervisionado para as tarefas com escassez de recursos, como a tarefa de RC. Nessa abordagem, através da engenharia de *prompt*, a tarefa a ser realizada é adaptada para o formato de tarefa que o LLM normalmente executa, dispensando assim a necessidade de aprendizado supervisionado e reduzindo significativamente a necessidade e o custo de recursos específicos (Liu *et al.*, 2021 *apud* Yang *et al.*, 2022).

Embora essa abordagem tenha sido implementada com êxito em várias tarefas de PLN, ela ainda é relativamente inexplorada no contexto da resolução de correferência (Sanh *et al.*, 2021 *apud* Yang *et al.*, 2022). No caso da língua portuguesa, nas pesquisas do presente trabalho não foram encontrados trabalhos publicados sobre o tema. Tal abordagem pode ser um recurso vantajoso para essa tarefa, especialmente considerando suas limitações de recursos - como a

escassez de grandes corpora para o treinamento supervisionado de modelos (sobretudo no português, que possui um volume ainda mais limitado de corpora anotados com informações de correferência (Fonseca *et al.*, 2018), quando comparado ao do inglês).

Exemplos de estudos que exploram essa possibilidade incluem os trabalhos de Sanh *et al.* (2021 *apud* Yang *et al.*, 2022), Le *et al*. (2022), Agrawa *et al*. (2022), Yang *et al.* (2022), Gan *et al.* (2024), Porada *et al*.(2024), Le *et al*. (2023) e Hicke *et al*. (2024).

O modelo T0, proposto por Sanh *et al.* (2021 *apud* Yang *et al.*, 2022), é uma generalização do modelo T5 (Raffel *et al.*, 2020 *apud* Yang *et al.*, 2022). Este modelo transforma tarefas específicas de PLN em *prompts* a partir de conjuntos de dados anotados, incluindo, entre outras, a resolução de correferência. Notavelmente, o modelo alcançou uma acurácia de mais de 60% no conjunto de dados *Winograd Challenge Scheme* (WSC) (Levesque *et al.*, 2012 *apud* Yang *et al.*, 2022). O WSC é um desafio de raciocínio de senso comum e compreensão de línguas naturais caracterizado por casos extremamente ambíguos dispostos em esquemas (*schemas)* - pares de sentenças que diferem em uma ou duas palavras com um mesmo pronome altamente ambíguo, cujo referente é diferente em cada uma das sentenças[22]. O WSC possui enfoque e formato específicos que o diferenciam de uma tarefa tradicional de RC (que envolvem a identificação e anotação de correntes de correferência em um corpus), o que torna os resultados obtidos menos indicativos da capacidade geral do modelo nesta área. No entanto, eles sugerem que esse tipo de modelo possa captar informações de correferência sem a necessidade de treinamento supervisionado específico para tal (Sanh *et al.*, 2021 *apud* Yang *et al.*, 2022). Outros modelos que exploram a RC em desafios no formato do WSC são os de Min *et al. (*2021 apud Anikina *et al., 2023*), Perez *et al*. (2021 apud Anikina *et al., 2023*) e Lin *et al*. (2023 apud Anikina *et al., 2023*).

Le *et al*. (2022) propõem o sistema MICE (*Mixtures of In-Context Expert*) para a resolução de anáfora em protocolos científicos com o uso de *prompts few-shot*. De acordo com os autores (tradução nossa):

> O MICE funciona da seguinte maneira: dada uma anáfora, como “a mistura”, o modelo utiliza aprendizado dentro do contexto em questão para predizer uma lista de substâncias contidas na mistura que sejam referenciadas anteriormente no procedimento descrito, como “brometo de bromoacetila”, “composto 54” e “água”.

[22] Exemplo de esquema:
Frase: "Os ativistas criticaram os grandes produtores rurais porque eles [promoviam/combatiam] a destruição das áreas de preservação."
Pergunta: Quem [promovia/combatia] a destruição das áreas de preservação?
Resposta: Os ativistas/os grandes produtores rurais

A partir de 16 ou 32 exemplos, o MICE gera um grupo de "especialistas contextuais", que consistem em *prompts* contendo algumas demonstrações selecionadas do desse conjunto de exemplos. As predições resultadas do uso dos especialistas são, então, combinadas em um modelo que as combina usando pesos calculados comparando os *embeddings* do *input* com as demonstrações codificadas por cada especialista. Dessa forma, cada especialista contribui com uma previsão ponderada com base na similaridade do *input* com as demonstrações que ele recebeu durante o treinamento. Isso permite ao MICE integrar diferentes perspectivas dos dados de treinamento para fornecer uma previsão final mais robusta e precisa sobre as substâncias mencionadas na anáfora.

Agrawa *et al*. (2022) utilizam *prompting* em LLMs - como o GPT-3 (Brown *et al.*, 2020) e o InstructGPT (Ouyang *et al.*, 2022 *apud* Agrawa *et al.*, 2022), uma versão do ChatGPT-3 ajustada por instruções (*instruction-tuned*)[23] - para várias tarefas de PLN na área clínica, incluindo a resolução de correferência. Os modelos são alimentados, por meio de *prompts*, com textos clínicos e uma menção – como, por exemplo, uma nota clínica e um pronome – e é solicitado ao modelo que identifique seu antecedente correto. São utilizados *prompts zero-shot* e *few-shot*. A abordagem simplifica o pós-processamento das respostas ao solicitar saídas estruturadas através de *prompts* personalizados, reduzindo a necessidade de processamento adicional complexo.

Yang *et al.* (2022), por sua vez, utilizam um método de *prompting* baseado em perguntas e respostas que solicita um *output* binário de sim ou não. Os experimentos foram conduzidos utilizando os modelos ChatGPT-2 e ChatGPT-Neo para a resolução do ECB+ (Cybulska *et al.*, 2014, *apud* Yang *et al.*, 2022), corpus com anotações de correferência intra e interdocumentais para entidades e eventos. O ChatGPT-2 também foi testado na resolução do WSC. A investigação centrou-se apenas na tarefa de avaliar a correferência entre pares de menções, sem abordar as demais subtarefas tradicionais da resolução de correferência: a detecção de menções e seu agrupamento em correntes de correferência. (Yang *et al.*, 2022)

A metodologia adotada por Yang *et al.* (2022) para a construção dos *prompts* envolveu fornecer um *input* contendo o texto e duas menções, seguido de uma pergunta sobre a correferência entre elas, solicitando uma resposta binária de sim ou não. Para os modelos *few-*

[23] Tipo de ajuste fino que visa aprimorar a capacidade de generalização de um modelo para novas tarefas, fazendo com que o modelo entenda e responda melhor a *prompts* de uma determinada área, mas sem treiná-lo para uma tarefa específica. (Zhang *et al.*, 2023)

shot[24], foram adicionados ao *prompt* exemplos no mesmo formato, com respostas já preenchidas. Observou-se que o modelo com *prompting* 0-shot apresentou uma baixa taxa de respostas dentro do formato esperado (5%), enquanto os modelos com 2, 4 e 10-shots alcançaram maior sucesso, com taxas de 93,7%, 96,2% e 98%, respectivamente. Os resultados foram avaliados com o uso das métricas de precisão, *recall*, F¹, acurácia e AUC[25], e podem ser verificados na Tabela 9.

Tabela 9 – Resultados obtidos por Yang *et al.* (2022)

Modelo	Corpus	Precisão	*Recall*	F¹	Acurácia	AUC
GPT-2	ECB+	0,08	0,53	0,14	0,50	0,51
GPT-NEO	ECB+	0,08	0,68	0,15	0,38	0,52
GPT-2	WSC	0,37	1,00	0,54	0,37	0,50

Fonte: Yang *et al.* (2022).

De acordo com Yang *et al.* (2022), os resultados obtidos indicam que o GPT-Neo realiza previsões que não são melhores do que resultados aleatórios, apesar de demonstrar consistência nas respostas para um mesmo *prompt*. Em contrapartida, o GPT-2 mostrou-se bastante inconsistente e sensível aos *prompts* utilizados. Comparados aos resultados do modelo e2e-coref (Lee *et al.*, 2017 *apud* Yang *et al.*, 2022), os resultados do GPT-Neo e GPT-2 foram notavelmente inferiores, especialmente devido à sua precisão extremamente baixa. Adicionalmente, foi constatado que os modelos testados possuem um desempenho relativamente melhor para resolução de correferência entre menções que incluam pronomes e entidades pessoais, bem como para pares de menções com alta similaridade. Essa observação sugere que, embora os modelos apresentem limitações significativas, há cenários específicos em que podem oferecer resultados mais satisfatórios. (Yang *et al.*, 2022).

Gan *et al.* (2024) avaliaram as capacidades dos modelos GPT-3.5, GPT-4 e de modelos da família LLAMA2 para a resolução de correferência. Eles criaram *prompts* 0-shot e *few-shot* baseados na arquitetura tradicional de *prompts* e na metodologia *Chain-of-Thought*[26] (Wei *et al.* 2023 *apud* Gan *et al.*, 2024). A pesquisa incluiu testes manuais e automáticos, com comparações entre ambos. Assim como em Yang *et al.* (2022), os modelos foram avaliados

[24] Modelos *few-shot* realizam tarefas a partir de *prompts* contendo poucos exemplos. Modelos *zero-shot* realizam tarefas sem o acesso a exemplos.
[25] Ver seção 2.6, "Métricas e *scorers*".
[26] Descrita na seção 3.2.

exclusivamente na subtarefa de reconhecimento de correferência entre pares de menções, sem abordar a etapa de identificação de menções ou de agrupamento de menções em correntes de correferência. Para isso, os pares de menções a serem analisados foram fornecidos nos *prompts*. (Gan *et al.*, 2024)

Os modelos foram testados em um documento da seção *Penn Treebank* do corpus ARRAU (Uryupina *et al.*, 2020 *apud* Gan *et al.*, 2024) das CRAC 2018[27]; um documento do subcorpus TRAIN, também do corpus ARRAU; e um documento do corpus LIGHT das CODI-CRAC 2021/2022. Além das versões genéricas do ChatGPT-3.5, ChatGPT-4 e LLAMA2-Chat 70B, também foram testados outros dois modelos da família LLAMA submetidos a ajustes finos utilizando o conjunto de treinamento das CRACs (Gan *et al.*, 2024). Os modelos genéricos também foram testados no Winogrande (Sakaguchi *et al.*, 2020 *apud* Gan *et al.*, 2024), conjunto de dados de grande escala com 44 mil problemas criado para avaliar o raciocínio de senso comum dos modelos de linguagem, onde o GPT-3.5 obteve 68% de precisão, o GPT-4 alcançou 94% e o LLama2 70B obteve 57% (Gan *et al.*, 2024).

Resultados da avaliação automática e manual dos modelos na resolução dos documentos das CRACs, nas métricas CoNLL e Universal Anaphora[28] podem ser conferidos na Tabela 10 e Tabela 11, respectivamente.

Tabela 10 – Resultados obtidos por Gan *et al.* (2024) em avaliação automática

Abordagem	LIGHT		Penn Treebank		TRAINS	
	UA	CoNLL	UA	CoNLL	UA	CoNLL
GPT3.5	54,44	32,18	49,05	25,34	42,55	31,63
GPT4	67,68	51,36	58,67	42,6	54,42	46,09
LLAMA2 70B	61,46	36,1	58	38,2	40,15	23,66
Baseline	75,62	67,48	83,9	76,48	65,82	58,16
LLAMA2 7B*	65,5	47,27	67,33	48,97	52,98	41,25
LLAMA 2 13B*	68,95	49,07	71,21	56,83	57,65	47,66

Legenda: "UA" representa a métrica Universal Anaphora. Os modelos submetidos a ajustes finos foram marcados com asterisco.
Fonte: Gan *et al.* (2024).

[27] CRACs as Competitions on Computational Models of Reference, Anaphora, and Coreference, tarefas compartilhadas que promovem o desenvolvimento e a avaliação de sistemas de resolução de correferência por meio de competições.
[28] Ver seção 2.6, "Métricas e *scorers*".

Tabela 11 – Resultados obtidos por Gan *et al.* (2024) em avaliação manual

Abordagem	LIGHT		Penn Treebank		TRAINS	
	UA	CoNLL	UA	CoNLL	UA	CoNLL
GPT4$_{(A1)}$	73,51	63,5	75,9	68,18	68,24	69,89
GPT$_{(A2)}$	77,59	69,45	78,35	72,1	75,05	74,08
LLAMA2 70B$_{(A1)}$	65,92	58,72	53,72	47,49	57,75	56,86
LLAMA2 70B$_{(A2)}$	70,91	61,85	58,11	50,71	62,6	62,41
LLAMA2 7B*	65,5	47,27	67,33	48,97	52,98	41,25
LLAMA 2 13B*	68,95	49,07	71,21	56,83	57,65	47,66

Legenda: $_{(A1)}$ e $_{(A2)}$ representam diferentes anotadores.
Fonte: Gan *et al.* (2024).

Com base na análise dos resultados obtidos por meio dos testes manuais e automáticos, Gan *et al.* (2024) concluíram que os métodos de avaliação vigentes não são plenamente adequados para mensurar a eficácia do uso de *prompting* em LLMs para a resolução de correferência, uma vez que tendem a subestimar injustamente o desempenho dos modelos, levando a uma classificação inferior à sua verdadeira capacidade.

Porada *et al.* (2024) propõem um método voltado especificamente para a tarefa de resolução de correferência pronominal que integra o uso de engenharia de *prompt* em um modelo de linguagem a um sistema supervisionado de RC. Considerando que o uso de *prompts* em LLMs para resolução de correferência (RC) demonstrou bom desempenho em problemas no estilo Winograd Schema Challenge (WSC), mas resultados inferiores em tarefas tradicionais de correferência, como as do CoNLL, a integração visa tirar proveito das vantagens de cada abordagem, criando um modelo mais competente na tarefa de RC pronominal.

Os autores adotam uma abordagem binária de pergunta e resposta, fornecendo ao modelo um trecho de texto com uma expressão pronominal e dois candidatos a antecedentes, para que o modelo determine qual dos dois candidatos possui relação de correferência com a expressão pronominal. Os testes envolvendo *prompts* se concentram em modelos da família Llama 3.1 em diferentes tamanhos, como 8B e 70B. O trabalho examina as versões com e sem ajustes finos desses modelos, integradas a um sistema supervisionado especializado em correferência, além de testar alguns modelos menores da série Llama 3.2. Os sistemas propostos foram testados utilizando 11 *datasets*, como o OntoNotes 5.0 (Wischedel *et al.*, 2013, apud

Porada *et al*., 2024) e o OntoGum (Zhu *et al*., 2021, apud Porada *et al*., 2024), além de conjuntos de problemas no estilo WSC, e avaliados por meio da métrica de acurácia.

Le *et al*. (2023) exploram a viabilidade de RC baseada em *prompts*, avaliando modelos de linguagem ajustados por instruções na resolução de tarefas compartilhadas de resolução de correferência, como o CoNLL-2012. Os formatos de *prompts* usados incluíram *prompts* de pergunta e resposta, em que o modelo recebe uma sentença e responde ao que se refere uma dada menção na sentença; e de *templates* de documento, em que o modelo é instruído a agrupar em correntes de correferência as menções pré-marcadas em um texto.

O estudo testou modelos da família Llama 2 (Touvron *et al*., 2023 apud Le *et al*., 2023), a saber: duas versões ajustadas por instruções do modelo da base Llama-2, Llama-2Chat e CodeLlama (Rozière *et al*., 2023). Também foram testados o InstructGPT, o ChatGPT (gpt-35-turbo) e ChatGPT-4. O sistema determinístico baseado em regras dcoref (Lee *et al*., 2013 apud Le *et al*., 2023) foi utilizado como linha de base.

Os resultados do formato de *prompt* de *templates* de documento foram avaliados utilizando o conjunto de teste da versão em inglês do OntoNotes 5.0 com menções previamente identificadas por duas fontes distintas: as menções douradas da anotação original do OntoNotes e menções preditas pelo dcoref. Os testes demonstraram que a qualidade da etapa de identificação de menções foi crucial para o desempenho dos modelos testados, que obtiveram desempenho significantemente melhor nos testes com menções douradas. Os resultados para as menções preditas e douradas podem ser verificados na Tabela 12 e 13, respectivamente.

Tabela 12 – Resultados obtidos por Le *et al.* (2023) em testes com menções preditas

Sistema	MUC	B^3	CEAFe	CoNLL
dcoref	67.7	55.9	52.5	58.6
weak-SpanBERT	68.6	56.7	52.7	59.3
Llama-2-Chat (70B)	39.7	42.3	22.2	34.7
CodeLlama (34B)	57.5	40.6	25.3	41.1
ChatGPT	66.9	55.5	46.5	56.3
InstructGPT	70.4	58.4	51.7	60.1
GPT-4	73.7	62.7	52.3	62.9

Fonte: Le *et al.* (2023).

Tabela 13 – Resultados obtidos por Le *et al.* (2023) em testes com menções douradas

Sistema	MUC	B^3	CEAFe	CoNLL
Menções douradas				
dcoref	81.6	70.0	67.3	72.9
Llama-2-Chat (7B)	19.7	40.2	22.8	27.6
Llama-2-Chat (70B)	58.2	65.7	34.4	52.8
CodeLlama (7B)	71.5	54.5	31.1	52.4
CodeLlama (34B)	75.6	66.5	43.1	61.7
ChatGPT	86.2	79.3	68.3	77.9
InstructGPT	89.2	79.4	73.7	80.8
GPT-4	93.7	88.8	82.8	88.4

Fonte: Le *et al.* (2023).

Alguns modelos também foram avaliados em outros conjuntos de dados para testar sua capacidade de generalização. Os testes envolveram corpora com escopo específico, como o Litbank (literatura) e WikiCoref (Wikipédia); subcorpora do OntoNotes com textos de diferentes períodos de tempo; e corpora de diferentes línguas, advindos dos conjuntos em chinês e árabe do OntoNotes e dos conjuntos de testes multilíngues do SemEval-2010. Como observa-se na Tabela 14, em testes com o fornecimento de menções douradas o InstructGPT teve bons resultados em línguas como chinês e holandês e resultados médios em línguas românicas como o catalão, espanhol e italiano.

Tabela 14 – Resultados do InstructGPT em corpora multilíngues

Língua	CoNLL *Score*
Chinês	77.3
Árabe	65.6
Catalão	41.9
Alemão	70.8
Italiano	41.4
Espanhol	42.2

Fonte: Le *et al.* (2023).

Hicke *et al.* (2024), por sua vez, propõem um modelo de RC que utiliza *prompting* em LLMs para gerar anotações de correferência em textos literários. O sistema proposto foca exclusivamente em anotações de correferência a nível de frase. Nesse modelo, o LLM recebe como *input* uma frase e a devolve anotada com informações de correferência. Como exemplo, o *input*:

1) "Carl enfiou suas mãos em seus bolsos, baixou sua cabeça, e disparou rua acima contra o vento norte."

Deve ser anotado pelo modelo como:

2) "[Carl: 1] enfiou [suas: 1] mãos em [seus: 1] bolsos, baixou [sua: 1] cabeça, e disparou [rua: 2] acima contra o vento norte."

O método utiliza dados de treinamento do corpus LitBank (Bamman *et al*., 2019 apud Hicke *et al*. 2024) e realiza ajustes finos em diferentes tamanhos de três LLMs: 4 tamanhos do T5 (Raffel *et al*., 2020 apud Hicke *et al*. 2024), 3 tamanhos do mT5 (Xue *et al*., 2021 apud Hicke *et al*. 2024) e 5 tamanhos do Pythia (Biderman *et al*., 2023; Hicke *et al*., 2024). Para analisar os resultados, utilizaram o CorefUD *scorer*, obtendo os resultados dispostos na Tabela 15.

Tabela 15 – Resultados obtidos por Hicke *et al.* (2024)

Modelo	MUC	B^3	CEAFm	CEAFe	BLANC	LEA	CoNLL
T5							
t5-3b	89.19	89.21	89.20	87.20	86.29	85.23	88.53
t5-large	82.14	83.71	83.41	81.19	77.81	77.74	82.35
t5-base	71.71	77.73	75.47	72.82	70.76	65.22	74.09
t5-small	45.62	55.82	52.03	48.10	41.97	36.47	49.85
mT5							
mT5-large	68.26	74.75	72.72	69.98	67.07	61.37	71.00
mT5-base	61.38	64.26	62.14	56.83	55.91	49.41	60.82
mT6-small	0.08	0.40	0.56	0.60	0.02	0.22	0.36

Fonte: autora.

Alguns outros trabalhos recentes exploram o uso de engenharia de *prompt* em modelos de linguagem para tarefas relacionadas à RC, embora de maneiras diferentes, atuando em funções que auxiliam ou estão envolvidas no processo de resolução, mas que não são a tarefa de resolução em si. O sistema DFKI-M*prompt*, por exemplo, elaborado por Anikina *et al.* (2023), foca na geração de menções utilizando *prompts* que solicitam a geração de possíveis menções e seus respectivos *spans*. Um sistema de resolução de correferências supervisionado usa, então, as menções geradas para estabelecer cadeias de correferência.

2.6 MÉTRICAS E *SCORERS*

Para as tarefas de resolução de correferência no formato de pares de menções, as métricas de F^1 e acurácia são as métricas padrão, por serem bastante intuitivas e amplamente usadas para tarefas de classificação binária. Entretanto, para a RC envolvendo o agrupamento de menções em correntes, a escolha de uma métrica de avaliação única torna-se mais complicada, uma vez que não há uma tarefa de classificação binária óbvia/única a ser avaliada por esse tipo de métrica nessas situações.

De acordo com Liu *et al.* (2023 – tradução nossa), "não há uma métrica única que seja universalmente apropriada para a resolução de correferência, devido à complexidade inerente à tarefa". As métricas podem abordar a tarefa de avaliação a partir de perspectivas distintas, concentrando-se em diferentes fatores. Em sua maioria, as diferentes métricas utilizadas na área de RC são, essencialmente, formas de aplicar as métricas tradicionais de precisão, *recall* e F^1 ao cenário mais complexo de agrupamento de menções. Como o formato da tarefa de RC envolvendo agrupamento de menções não é perfeitamente adequado ao uso dessas métricas tradicionais, é necessário fazer uma decisão sobre o quê, exatamente, será avaliado em termos de positivo, negativo, verdadeiro e falso. Para isso, algumas métricas adotam uma abordagem **baseada em *links***, que avalia as conexões entre menções, enquanto outras seguem uma abordagem **baseada em entidades**, que considera a associação das menções a cada grupo (Sukthanker *et al.*, 2020). Diferentes métricas exploram diferentes maneiras de fazê-lo, buscando mitigar diferentes desafios encontrados na avaliação dos sistemas. Por conta disso, considerando as vantagens e desvantagens de cada métrica desenvolvida, frequentemente se utiliza uma combinação de métricas que contemplem diferentes aspectos da avaliação de sistemas de RC. (Liu *et al.*, 2023)

2.6.1 F¹

Precisão é uma métrica que divide o número de verdadeiros positivos em relação ao total de positivos previstos (verdadeiros e falsos). Ou seja, mede a exatidão dos resultados positivos identificados:

$$Precisão = \frac{PositivosVerdadeiros}{PositivosVerdadeiros + FalsosPositivos}$$

Recall é a proporção de verdadeiros positivos identificados em relação ao total de casos positivos reais, medindo a capacidade do modelo em não deixar de detectar positivos:

$$Recall = \frac{PositivosVerdadeiros}{PositivosVerdadeiros + FalsosNegativos}$$

O **F¹** é a média harmônica da precisão e *recall*, proporcionando uma medida equilibrada que considera tanto a exatidão quanto a cobertura dos resultados positivos:

$$F^1Score = \frac{2 * Precisão * Recall}{Precisão + Recall}$$

Há também a medida de acurácia, que é a proporção de predições corretas (verdadeiros positivos e verdadeiros negativos) em relação ao número total de predições, sendo uma métrica geral de correção:

$$Acurária = \frac{VerdadeirosPositivos + VerdadeirosNegativos}{TotaldePredições}$$

2.6.2 MUC

Segundo Liu *et al.* (2023) essa métrica, introduzida por Vilain *et al.* (1995, *apud* Liu *et al.*, 2023), baseia-se no número de *links* - a relação de correferência entre duas menções - corretamente identificados. Essencialmente, o MUC atribui uma pontuação baseada no número

de modificações de *links* necessários para fazer o conjunto de resposta idêntico ao conjunto de referência (Sukthanker *et al.*, 2020).

Na métrica MUC, a precisão avalia a exatidão dos *links* de correferência previstos por um sistema, concentrando-se na proporção de *links* corretos dentre aqueles identificados. Ela mede o quão bem o sistema evita divisões desnecessárias entre menções que deveriam estar agrupadas. Uma maior precisão indica que a maioria dos *links* previstos corresponde com precisão aos verdadeiros agrupamentos presentes no conjunto de referência. (Sukthanker *et al.*, 2020)

Já o *recall* mede a capacidade do sistema de recuperar os *links* de correferência verdadeiros. Ele calcula a proporção de *links* corretos do conjunto de referência que o sistema conseguiu identificar. Um *recall* mais alto sugere que o sistema foi eficaz em capturar a maioria dos *links* verdadeiros, garantindo que menções que se referem à mesma entidade fossem agrupadas corretamente. (Sukthanker *et al.*, 2020)

2.6.3 B³ (B-Cubed)

Formulada por Bagga e Baldwin (2019), a métrica B³ avalia a precisão e o *recall* com base em cada menção, combinando-os em uma média ponderada (Liu *et al.*, 2023). A métrica atribui um peso para cada menção individual, baseando-se na quantidade de menções da corrente a qual faz parte ou foi atribuída e o total de menções existentes (Fonseca, 2018; Sukthanker *et al.*, 2020).

Para calcular a precisão e o *recall* de cada menção *i*, segue-se a lógica abaixo:

$$Precisão_i = \frac{nº\ de\ elementos\ corretos\ na\ corrente\ respondida\ à\ qual\ foi\ associada}{número\ de\ elementos\ totais\ na\ corrente\ respondida\ à\ qual\ foi\ associada}$$

$$Recall_i = \frac{número\ de\ elementos\ corretos\ na\ corrente\ respondida\ à\ qual\ foi\ associada}{número\ de\ elementos\ totais\ na\ corrente\ verdadeira\ à\ qual\ foi\ associada}$$

Os índices de precisão e *recall* do sistema são, então, calculados com base na soma ponderada da precisão e *recall* de cada menção individual. (Sukthanker *et al.*, 2020)

2.6.4 CEAF

A métrica CEAF, proposta por Luo (2005, *apud* Liu *et al.*, 2023), avalia a resolução de correferência baseada no alinhamento de menções e entidades a fim de detectar sua similaridade. Faz isso criando um mapeamento alinhando as correntes de correferência respondidas pelo modelo avaliado às correntes originais - a partir de opções de fórmulas de alinhamento de maior ou menor exigência -, e então calcula sua acurácia e *recall* a partir deste mapeamento. Possui duas variações: CEAFm (baseada em menções) e CEAFe (baseada em entidades). (Liu *et al.* 2023)

2.6.5 BLANC

Diferente das métricas tradicionais de avaliação de correferência, a métrica BLANC, introduzida por Recasens *et al.* (2011, *apud* Liu *et al.*, 2023), considera tanto os *links* de correferência quanto os de não-correferência. Desta maneira, consegue abordar os *singletons* - entidades que aparecem apenas uma vez em um texto -, que geralmente são negligenciadas por outras métricas. (Liu *et al.* 2023)

A avaliação pelo BLANC envolve dois tipos principais de elementos: *links* de correferência, que denotam pares de menções que se referem à mesma entidade, e *links* de não-correferência, indicando pares que não se referem à mesma entidade. Desta maneira, atribui um acerto quando o sistema original e o sistema de resposta atribuem o mesmo valor (correferência ou não-correferência) a um *link*, e um erro quando não o fazem. A métrica mede a precisão e o *recall* para ambos os tipos de *links* e calcula duas médias separadas: uma média das precisões e outra média dos *recall*s. (Recasens e Hovy, 2011, *apud* Liu *et al.*, 2023)

$$Precisão^{c} = \frac{número\ de\ links\ de\ correferência\ corretamente\ atribuídos}{número\ de\ links\ de\ correferência\ atribuídos}$$

$$Recall^{c} = \frac{número\ de\ links\ de\ correferência\ corretamente\ atribuídos}{número\ de\ links\ de\ correferência\ reais}$$

$$Precisão^{n} = \frac{número\ de\ links\ de\ não-correferência\ corretamente\ atribuídos}{número\ de\ links\ de\ não-correferência\ atribuídos}$$

$$Recall^{n} = \frac{número\ de\ links\ de\ não - correferência\ corretamente\ atribuídos}{número\ de\ links\ de\ não - correferência\ reais}$$

$$Precisão = \frac{Precisão^{c} + Precisão^{n}}{2}$$

$$Recall = \frac{Recall^{c} + Recall^{n}}{2}$$

De acordo com Recasens *et al.* (2011, *apud* Liu *et al.*, 2023), o design da métrica BLANC permite lidar com casos em que o número de entidades ou a natureza das menções de entidades diferem significativamente entre o padrão de ouro e a saída do sistema. Ela faz isso mantendo um número constante total de *links* (tanto de correferência quanto de não-correferência) em ambos os conjuntos, garantindo que a avaliação não seja enviesada pela tendência do sistema de prever excessivamente ou subestimar o número de entidades. (Recasens *et al.*, 2011, *apud* Liu *et al.*, 2023)

2.6.6 LEA

LEA (Moosavi *et al.*, 2016, *apud* Liu *et al.*, 2023) é uma métrica que calcula uma pontuação baseada na importância das entidades do documento original e em quão bem elas são resolvidas. A importância de uma entidade é medida pelo número de menções que ela contém, de forma que acertos em entidades com mais menções contribuem mais para a pontuação final. A métrica também pode ser adaptada para considerar outros fatores, como o tipo de entidade ou o tipo de menção, dependendo da tarefa final ou do domínio em questão. (Sukthanker *et al.*, 2020)

O *recall* é calculado ponderando a importância de cada entidade no conjunto de referência, definida pelo seu tamanho, multiplicada pelo número de *links* corretamente atribuídos entre suas menções. Em seguida, essa soma é dividida pela importância somada de todas entidades no conjunto de referência. (Sukthanker *et al.*, 2020)

A precisão é calculada ponderando a importância de cada entidade no conjunto de respostas multiplicada pelo número de *links* corretamente atribuídos entre suas menções. Em seguida, essa soma é dividida pela importância somada de todas as entidades no conjunto de respostas. (Sukthanker *et al.*, 2020)

2.6.7 Combinação de métricas e *scorers*

Como mencionado anteriormente, a utilização de combinações de métricas é uma prática comum na avaliação de sistemas de RC. A combinação mais frequente, adotada pelos modelos mais relevantes nos últimos anos, é a abordagem utilizada na tarefa compartilhada CoNLL 2012 (Pradhan *et al*., 2012). Conhecida também como *CoNLL Score*, consiste na média dos valores de F^1 das métricas B^3, MUC e CEAF (Liu *et al*., 2023).

Para automatizar essa combinação de métricas, existem algumas ferramentas, como o CorefUD *scorer*[29] (Github, 2024) e o Universal Anaphora (UA) *Scorer*[30] (Yu *et al.*, 2022; Yu *et al*., 2023). O UA *Scorer* é o avaliador da Universal Anaphora[31]. Sua versão 1.0 suporta o formato de anotação da iniciativa UA e se baseia no Reference Coreference *Scorer* (Pradhan *et al.*, 2014; Strube *et al.*, 2016; Poesio *et al.*, 2018 apud Yu *et al.*, 2022), utilizado nas tarefas compartilhadas CoNLL-2011/2012 em resolução de correferência. O UA *scorer* se destaca por suportar a avaliação de casos de anáfora indireta e anáfora com antecedentes coordenados.

O CorefUD, por sua vez, é o avaliador oficial para as tarefas compartilhadas das CRACs em resolução de correferência multilíngue. A ferramenta suporta anotações no formato CorefUD 1.0 e trabalha com correspondências exatas, parciais e de núcleo das menções, abrangendo todas as principais métricas usadas para correferência, como B^3, CEAF, BLANC e LEA. O avaliador é limitado à avaliação da resolução de correferências, não contemplando casos de antecedentes coordenados, anáfora indireta e outras relações, e é construído com base no Universal Anaphora *scorer* 1.0 (GitHub, 2024). Mais tarde, Yu *et al*. (2023) basearam-se no CorefUD para expandir a funcionalidade do UA *Scorer*, incorporando o suporte ao formato CorefUD 1.0 na versão UA *scorer* 2.0.

---

[29] Disponível para acesso em: https://github.com/ufal/corefud-*scorer*

[30] Disponível para acesso em: https://github.com/juntaoy/universal-anaphora-*scorer*?tab=readme-ov-file

[31] Iniciativa que coleta e distribui recursos para a resolução de anáforas e correferência com base em diretrizes e formato unificados, visando expandir o escopo de informações de correferência /anáfora anotáveis de forma consistente em corpora (Yu *et al.*, 2022).

3 METODOLOGIA

3.1 CORPUS DE TESTAGEM

3.1.1 Escolha do corpus

Para o presente trabalho, optou-se pela utilização de um fragmento de um corpus com anotações de correferência do português já existente, porém adaptado a um formato mais atualizado que possibilita o acesso a uma gama maior e mais recente de recursos, como *scorers*, *parsers* e outras ferramentas que facilitam a manipulação e análise dos dados. Para essa adaptação, foi escolhido o formato da iniciativa CorefUD (Nedoluzhko *et al.*, 2022), que visa a padronização de anotações de correferência em corpora de diferentes línguas.

Embora o Corref-PT tenha sido inicialmente escolhido, por ser o maior e mais recente corpus anotado com informações de correferência no português, algumas inconsistências na formatação do texto dificultaram sua conversão para o formato desejado. Como exemplo, destacam-se ocorrências da inclusão simultânea de sintagmas nominais e seus sintagmas embutidos na mesma corrente de correferência, como “Sons of Anarchy Motorcycle Club, Redwood Original”, “Anarchy Motorcycle Club” e “Redwood Original”, ou “A seleção brasileira masculina de vôlei” e “vôlei”, na Figura 8.

Figura 8 – Segmento do corpus Corref-PT

ID	Token	Lemma	PoS	Feat	Head	Coref
23	(	(	(	-	_	_
24	Sons		prop	M=S	0	(0
25	of		,	-	_	_
26	Anarchy_Motorcycle_Club		prop	M=S	_	0
27	,		,	-	_	_
28	Redwood_Original		prop	M=S	_	0)|(0)
[...]						
1	A	o	art	F=S	_	(0
2	seleção	seleção	n	F=S	0	_
3	brasileira	brasileiro	adj	F=S	_	_
4	masculina	masculino	adj	F=S	_	_
5	de	de	prp	-	_	_
6	vôlei		n	M=S	_	0)|(0)

Fonte: autora.

No mesmo exemplo, também nota-se uma escolha inconsistente na anotação de entidades multipalavras, como é o caso de "Anarchy_Motorcycle_Club", que faz parte do nome próprio *Sons of Anarchy Motorcycle Club*, mas foi marcada separadamente do restante do nome. Há, também, inconsistência na indexação dos *tokens* - na primeira sentença de cada texto, utiliza-se uma indexação de base 0, mas utiliza-se uma indexação de base 1 para as demais sentenças.

Devido a limitações de tempo e recursos e ao fato de que o objetivo deste trabalho não requer um corpus extenso, uma vez que não haverá etapas de treino e validação, foi utilizado, em seu lugar, o segundo maior corpus atual do português com informações de correferência entre sintagmas nominais[32], o Summ-It++.

3.1.2 CorefUD

CorefUD (Nedoluzhko, 2022) se trata de uma coleção multilíngue de corpora anotados com informações de correferência que visa mitigar a falta de padronização nos formatos de dados e nas diretrizes desse tipo de anotação. Para isso, harmoniza as anotações dos corpora de maneira compatível com o framework das *Universal Dependencies* (Marneffe *et al.*, 2021), projeto que fornece um inventário universal de categorias e diretrizes para facilitar e padronizar a anotação morfossintática de textos em diferentes línguas, criando uma coleção multilíngue de corpora anotados de forma padronizada, com o objetivo de facilitar o desenvolvimento de *parsers* e outros sistemas e aplicações multilíngues de PLN (Marneffe *et al.*, 2021).

As *Universal Dependencies* utilizam o formato CoNLL-U[33], extensão amplamente utilizada para a anotação de dependências sintáticas em linguística computacional. O CoNLL-U segue um formato tabular com colunas predefinidas que incluem informações sobre tokens, como identificador de palavra (ID), forma (FORM), lema (LEMMA), *tag* universal de parte do discurso (UPOS), tag opcional morfológico ou de parte do discurso de uma língua específica (XPOS), lista de características morfológicas (FEATS), núcleo da palavra atual (HEAD), relação de dependência universal (DEPREL), grafo de dependência aprimorado (DEPS), e anotações com quaisquer outras informações que não se encaixem nas categorias padrão

[32]Apesar de o corpus de Garcia *et al.* (2014) ter um maior número de tokens, este trata apenas das relações de correferência entre entidades pessoais, deixando de fora as demais relações entre sintagmas nominais.
[33] Mais detalhes em: https://universaldependencies.org/format.html

(MISC). Cada linha representa um *token* (geralmente uma palavra, mas podem haver tokens multipalavra), e as colunas são separadas por tabulações. (CoNLL-U format, 2024).

O CorefUD adota o mesmo formato e o mantém intacto ao aproveitar coluna MISC para a anotação dos atributos de correferência, evitando a adição de uma nova coluna para a finalidade. Dessa forma, garante a conformidade aos os padrões das *Universal Dependencies*, que exigem o uso do formato CONLL-U sem qualquer tipo de alteração. (Nedoluzhko, 2022)

O formato CorefUD utiliza uma estrutura de atributos adicionais, como "Entity" para identificar menções, tipo de entidade, núcleo da menção e tipo de correferência; "Bridge" para indicar casos de anáfora indireta e “SplitAnt” para indicar casos de anáforas com antecedentes coordenados. Cada menção é especificada por um *span*, ou seja, o conjunto de palavras que ela abrange, e cada menção pertence a um único *cluster* que representa uma entidade ou evento. As menções podem ser contínuas ou descontínuas, e a anotação permite a inclusão de menções que se estendem por múltiplas sentenças. (Nedoluzhko, 2022)

Figura 9 – Exemplo de trecho do corpus Summ-it++ na formatação CoNLL-U com informações de correferência na formatação CorefUD

newdoc id = CIENCIA_2001_19858
global.Entity = eid-etype-head-other
sent_id = s1
text = Cientistas de o Centro de Estudos de Saclay, em a França, parecem ter encontrado o primeiro indício de que a Via Láctea esteja cercada por um verdadeiro campo minado de buracos negros, cerca de 1 milhão a o todo, eles estimam.

ID	FORM	LEMMA	UPOS	XPOS	FEATS	HEAD	DEPREL	DEPS	MISC
1	Cientistas	cientista	NOUN	_	Number=Plur	14	nsubj	_	Entity=(e1-person-1-coref
2	de	de	ADP	_	_	4	case	_	_
3	o	o	DET	_	Definite=Def|Gender=Masc|Number=Sing|PronType=Art	4	det	_	_
4	Centro	Centro	PROPN	_	Gender=Masc|Number=Sing	1	nmod	_	_
5	de	de	PROPN	_	_	4	flat	_	_
6	Estudos	Estudos	PROPN	_	_	4	flat	_	_
7	de	de	PROPN	_	_	4	flat	_	_
8	Saclay	Saclay	PROPN	_	_	4	flat	_	Entity=e1)|SpaceAfter=No
9	,	,	PUNCT	_	_	12	punct	_	_

Fonte: autora.

Após contato com a organização da iniciativa CorefUD a respeito da harmonização do fragmento do corpus Summ-it++ e com a perspectiva de, posteriormente, harmonizar todo o corpus para integrá-lo aos próximos lançamentos do CorefUD, foi tomada a decisão de fazer uma divisão do corpus em conjuntos de treinamento, desenvolvimento e testagem de proporções 80%, 10% e 10%, respectivamente, dada a inexistência de uma divisão prévia. Optou-se por trabalhar, inicialmente, apenas com o conjunto de desenvolvimento. Essa opção se deve ao fato de que, para a publicação do corpus na iniciativa, é recomendável que o conjunto

de teste seja inicialmente mantido em sigilo. Os corpora publicados nos lançamentos do CorefUD são normalmente utilizados nas CRACs. Nessas competições, os corpora são divididos em conjuntos de treinamento, validação e teste, sendo o conjunto de teste mantido em sigilo até a fase final da competição para assegurar uma avaliação justa e imparcial dos sistemas participantes. Como os testes do presente trabalho foram realizados em um LLM com o objetivo de avaliar sua capacidade generalizada de resolução de correferência, de forma que não haja a necessidade de um treinamento com um corpus anotado, não se viu problema em dispensar o conjunto de treinamento e utilizar o conjunto de desenvolvimento como conjunto de teste.

Portanto, para a finalidade deste estudo, serão empregados o conjunto de desenvolvimento da versão harmonizada e revisada do corpus Summ-it++, correspondendo a 10% da totalidade do corpus original; e a versão apenas harmonizada, mas não revisada, deste mesmo segmento do corpus.[34] O Summ-it++ está publicado sob a licença *Creative Commons Attribution-NonCommercial 4.0 International*, que permite que seja compartilhado, adaptado e utilizado para a elaboração de outros materiais, desde que sejam atribuídos os devidos créditos aos autores originais, mas não para uso comercial.

3.1.3 Harmonização do corpus

Para a harmonização do segmento do corpus Summ-it++, foi inicialmente refeita a anotação morfossintática do corpus, de forma automática, utilizando o UDPipe (Straka *et al.*, 2016), a fim de adequá-la às diretrizes das Universal Dependencies. O UDPipe é uma ferramenta que realiza segmentação de sentenças, tokenização, anotação de partes do discurso), lematização e *parsing* de dependências. A ferramenta suporta mais de 50 idiomas, dentre eles o português. (Straka *et al.*, 2016)

No corpus original, o campo "Head" denota se a palavra é o núcleo de um sintagma nominal, caso no qual o campo recebe o valor '0'. Para todas as outras palavras, o campo recebe o valor "_" (Antonitsch *et al.*, 2016). O campo HEAD no formato CoNLL-U, por sua vez, é utilizado para indicar a palavra núcleo de cada dependência em uma sentença anotada. Para cada palavra, esse campo contém o índice da palavra à qual a palavra atual está conectada como dependente. (CoNLL-U format, 2024).

[34] Corpora disponíveis em: https://github.com/manupac/Summ-it-CorefUD

O Udapi[35], framework em Python para processamento de dados das *Universal Dependencies*, dispõe de um bloco, denominado *MoveHead*, dedicado à atribuição automática dos núcleos das menções das correferências anotadas com base na árvore de dependências do corpus, caso o sistema a ser testado não possua a função de atribuí-las ele mesmo (como é o caso do uso dos modelos no presente trabalho).

A anotação do campo "Head" do corpus original revelou-se inadequada para inferir automaticamente qual palavra é o núcleo de uma menção, uma vez que pode haver múltiplos núcleos de sintagmas nominais dentro de um mesmo intervalo de menção e a estrutura completa de dependências da frase, que permitiria essa inferência, não é anotada nesse formato. Em virtude dessa limitação, essa anotação foi descartada e, em seu lugar, implementou-se a anotação completa de dependências fornecida pelo UDPipe.

Além disso, apesar de o corpus original conter informações sobre tipos de entidade, estas não puderam ser incorporadas à sua versão harmonizada. Isso se deve à ausência dessas informações na maioria das menções (como observa-se nos trechos da Figura 10 à Figura 16, na Seção 3.1.3.1) e ao fato de que, nas menções em que estão presentes, os *spans* das anotações de tipos de entidade e das anotações de correferência nem sempre coincidem, como ocorre no exemplo da Figura 17 (Seção 3.1.3.1), em que o *span* da anotação sobre o tipo de entidade (PER, para “pessoa”, na coluna 7) vai do token 38 ao 39, enquanto que o *span* da anotação da menção abrange apenas o token 39.

Após a conclusão das adequações de formato e anotações morfossintáticas do corpus, foi feita uma revisão das informações de correferência.

3.1.3.1 Revisão de anotações de correferência

Além da harmonização do formato das anotações de correferência do Summ-it++ ao formato CorefUD, também foi realizada uma revisão manual dessas anotações. A maioria das anotações originais foi mantida, porém foram corrigidas algumas inconsistências identificadas. Exemplos dessas inconsistências no corpus original incluem:

a) Menções atribuídas simultaneamente a entidades diferentes - como a menção "um buraco negro", no exemplo da Figura 10, que é anotada como pertencente, ao mesmo tempo, às correntes de correferência 28 e 29.

[35] Pode ser acessado em: https://github.com/udapi/udapi-python/tree/master

Figura 10 – Segmento do corpus Summ-it++

ID	Token	Lemma	PoS	Feat	Head	NE	Rel	Coref
26	um	um	art	M=S	_	_	_	(28|(29
27	buraco	buraco	n	M=S	0	_	_	_
28	negro	negro	adj	M=S	_	_	_	28)|29)

Fonte: autora.

b) Menções não atribuídas à corrente de correferência à qual deveriam pertencer - como no caso do trecho "o disco galáctico", no exemplo da Figura 11, que em sua primeira ocorrência é atribuído à corrente de correferência 4, mas que não é atribuído a nenhuma corrente em sua segunda ocorrência no mesmo texto.

Figura 11 – Segmento do corpus Summ-it++

ID	Token	Lemma	PoS	Feat	Head	NE	Rel	Coref
31	o	o	art	M=S	_	_	_	_
32	centro	centro	n	M=S	0	_	_	_
33	de	de	prp	_	_	_	_	_
34	a	o	art	F=S	_	_	_	(7
35	Via_Láctea	_	prop	F=S	0	(PLC)	_	7)
36	fora_do	_	prp	_	_	_	_	_
37	o	o	art	M=S	_	_	_	_
38	plano	plano	n	M=S	0	_	_	_
39	de	de	prp	_	_	_	_	_
40	o	o	art	M=S	_	_	_	(4
41	disco	disco	n	M=S	0	_	_	_
42	galáctico	galáctico	adj	M=S	_	_	_	4)
43	.	.	.	_	_	_	_	_
[...]								
39	ele	ele	pron-pers	M=3S=NOM	0	_	_	(29)
40	não	não	adv	_	_	_	_	_
41	pertencia	pertencer	v-fin	IMPF=1S=IND	_	_	_	_
42	a	a	prp	_	_	_	_	_
43	o	o	art	M=S	_	_	_	_
44	disco	disco	n	M=S	0	_	_	_
45	galáctico	galáctico	adj	M=S	_	_	_	_
46	.	.	.	_	_	_	_	_

Fonte: autora.

c) Entidades mencionadas para as quais não são criadas correntes de correferência - como é o caso da estrela anã mencionada nos segmentos das Figuras 12 e 13, que é referida

pelas menções "uma estrela anã" e "a estrela vizinha", respectivamente, mas à qual não é atribuída nenhuma corrente de correferência, de forma que suas menções não sejam anotadas.

Figura 12 – Segmento do corpus Summ-it++

ID	Token	Lemma	PoS	Feat	Head	NE	Rel	Coref
0	O	o	art	M=S	_	_	_	(29
1	objeto	objeto	n	M=S	0	_	_	_
2	em	em	prp	_	_	_	_	_
3	questão	questão	n	F=S	0	_	_	_
4	(	(	(	_	_	_	_	_
5	um	um	art	M=S	_	_	_	_
6	sistema	sistema	n	M=S	0	_	_	_
7	duplo	duplo	adj	M=S	_	_	_	29)
8	composto	compor	v-pcp	M=S	_	_	_	_
9	por	por	prp	_	_	_	_	_
10	um	um	art	M=S	_	_	_	_
11	buraco	buraco	n	M=S	0	_	_	_
12	negro	negro	adj	M=S	_	_	_	_
13	sendo	ser	v-ger	_	_	_	_	_
14	orbitado	orbitar	v-pcp	M=S	_	_	_	_
15	por	por	prp	_	_	_	_	_
16	uma	um	art	F=S	_	_	_	_
17	estrela	estrela	n	F=S	0	_	_	_
18	anã	anão	n	F=S	0	_	_	_
19	)	)	)	_	_	_	_	_

Fonte: autora.

Figura 13 – Segmento do corpus Summ-it++

ID	Token	Lemma	PoS	Feat	Head	NE	Rel	Coref
10	o	o	art	M=S	_	_	_	_
11	quanto_a	_	prp	_	_	_	_	_
12	estrela	estrela	n	F=S	0	_	_	_
13	vizinha	vizinho	n	F=S	0	_	_	_
14	foi	ir, ser	v-fin	PS=3S=IND	_	_	_	_
15	consumida	consumir	v-pcp	F=S	_	_	_	_
16	por	por	prp	_	_	_	_	_
17	o	o	art	M=S	_	_	_	(29
18	buraco	buraco	n	M=S	0	_	_	_
19	negro	negro	adj	M=S	_	_	_	29)

Fonte: autora.

d) Inconsistências na anotação de casos de anáfora indireta (ou *bridging*). A menção "O objeto em questão (um sistema duplo).", no trecho da Figura 12, que se refere ao sistema composto por um dado buraco negro e a estrela anã que o cerca; e a menção "o buraco negro", no trecho da Figura 13, que se refere a esse mesmo buraco negro;

são ambas atribuídas à mesma corrente de correferência. Outras situações análogas, contudo, são tratadas de maneira diferente, como a situação do exemplo da Figura 14, em que a menção ao cientista Felix Mirabel não é associada à menção ao grupo de cientistas do qual faz parte.

Figura 14 – Segmento do corpus Summ-it++

ID	Token	Lemma	PoS	Feat	Head	NE	Rel	Coref
19	a	o	art	F=S	_	_	_	_
20	Folha_Felix_Mirabel	_	prop	F=S	0	_	_	_
21	,	_	,	_	_	_	_	_
22	pesquisador	pesquisador, pesquisar	n	M=S	0	_	_	_
23	que	que	pron-indp	M=S	_	_	_	_
24	liderou	liderar	v-fin	PS=3S=IND	_	_	_	_
25	o	o	art	M=S	_	_	_	(22
26	grupo	grupo	n	M=S	0	_	_	22)
27	.	.	.	_	_	_	_	_

Fonte: autora.

e) Inconsistências nas anotações dos casos de antecedentes coordenados. Há diversas ocorrências em que uma anáfora com antecedentes coordenados é associada apenas a um dos seus antecedentes, e não ao(s) outro(s), como é o caso do exemplo da Figura 15 em que a menção “eles” é associada a “neandertais”, mas não a “humanos modernos”).

Figura 15 – Segmento do corpus Summ-it++

ID	Token	Lemma	PoS	Feat	Head	NE	Rel	Coref
12	investigar	investigar	v-inf	_	_	_	_	_
13	a	o	art	F=S	_	_	_	_
14	relação	relação	n	F=S	0	_	_	_
15	entre	entre	prp	_	_	_	_	_
16	neandertais	_	n	M=P	0	_	_	(2)
17	e	e	conj-c	_	_	_	_	_
18	humanos	humanos	n	M=P	0	_	_	_
19	modernos	moderno	adj	M=P	_	_	_	_
20	olhando	olhar	v-ger	_	_	_	_	_
21	não	não	adv	_	_	_	_	_
22	para	para	prp	_	_	_	_	_
23	seus	seu	pron-det	M=P	0	_	_	_
24	crânios	crânio	n	M=P	_	_	_	_
25	,	_	,	_	_	_	_	_
26	mas	mas	conj-c	_	_	_	_	_
27	para	para	prp	_	_	_	_	_
28	o	_	pron-det	M=S	_	_	_	_
29	que	que	pron-indp	M=S	0	_	_	_
30	eles	ele	pron-pers	M=3P=NOM	0	_	_	(2)
31	defecavam	defecar	v-fin	IMPF=3P=IND	_	_	_	_

Fonte: autora.

f) Menções cujo *span* engloba trechos que não fazem parte da menção, como no exemplo da Figura 16, em que a menção anotada é “fezes possível”, em que “possível” se refere à “quantidade de fezes” e não deve, portanto, ser associado ao substantivo “fezes”.

Figura 16 – Segmento do corpus Summ-it++

ID	Token	Lemma	PoS	Feat	Head	NE	Rel	Coref
17	a	o	art	F=S	_	_	_	_
18	maior	maior	adj	F=S	_	_	_	_
19	quantidade	quantidade	n	F=S	0	_	_	_
20	de	de	prp	_	_	_	_	_
21	fezes	fezes	n	F=P	0	_	_	(3
22	possível	possível	adj	M=S	_	_	_	3)

Fonte: autora.

Figura 17 – Segmento do corpus Summ-it++

ID	Token	Lemma	PoS	Feat	Head	NE	Rel	Coref
2	sir	_	v-inf	_	_	(PER	_	_
3	David_King	_	prop	M=S	0	PER)	_	(11)

Fonte: autora.

Conforme Mitkov (2002, p. 141, tradução nossa), “quando comparada à análise sintática, a análise de relações anafóricas a nível de discurso envolve um processo muito mais interpretativo, e a possibilidade de desacordo de interpretação entre anotadores é muito maior”. Segundo o autor, a complexidade das tarefas de anotação de anáfora e de correferência impõe uma decisão entre mais simplicidade na natureza das anotações, de forma a possibilitar mais confiabilidade e concordância entre anotadores; ou anotações mais completas e detalhadas, porém sujeitas à um maior grau de subjetividade e menos utilizáveis no campo do PLN. Analisando a Tabela 16, elaborada por Liu *et al.* (2023), nota-se que a maioria dos corpora mais relevantes do inglês se restringem a informações de correferência de entidade, deixando de fora as informações sobre correferência de evento.

Tabela 16 – Escopo dos principais corpora anotados com informação de correferência

(continua)

Corpus	Escopo	
	Entidade	Evento
CoNLL 2012	✓	
GAP	✓	

Tabela 16 – Escopo dos principais corpora anotados com informação de correferência (conclusão)

Corpus	Escopo	
	Entidade	Evento
CoNLL 2012	✓	
GAP	✓	
KBP 2017		✓
ACE 2005		✓
LitBank	✓	
WSC	✓	
DPR	✓	
PDP	✓	
Winogender	✓	
WinoBias	✓	
KnowRef	✓	
WikiCoref	✓	
ECB+	✓	✓
Red		✓
GUM	✓	
WEC		✓
EmailCoref	✓	
BUG	✓	

Fonte: Liu *et al.* (2023).

Além disso, mesmo em corpora com anotação somente de correferências entre sintagmas nominais, o processo de decidir o que deve ser marcado como correferente não é simples (Mitkov, 2002). De acordo com Mitkov *et al.* (2002, p.142, tradução nossa):

> Uma estratégia de anotação em forma de diretrizes delimitando o que anotar e recomendando as melhores práticas de anotação pode, além de ser muito útil para os anotadores, melhorar a consistência das anotações e o acordo entre anotadores, que são frequentemente baixos. (Mitkov, 2002, p. 142, tradução nossa).

Dessa forma, para fins de padronização e clareza na revisão das anotações de correferência do corpus do presente trabalho, foram mantidas, excluídas ou adicionadas anotações à versão original do Summ-it++ com base nas diretrizes de anotação de correferência no inglês da versão 5.0 do corpus OntoNotes[36], corpus amplamente utilizado em pesquisas sobre correferência no campo do PLN em inglês, com algumas adaptações/alterações (Pradhan *et al.*, 2012). Conforme Pradhan *et al.* (2012), essas diretrizes são consideravelmente independentes de uma língua específica e podem ser aplicadas a diferentes idiomas.

3.1.3.2 Diretrizes do OntoNotes

O conjunto de diretrizes de anotação do OntoNotes abrange a anotação tanto de correferência de entidades quanto de eventos. As correferências são classificadas em dois tipos principais: de identidade e apositivas, que são anotados separadamente, pois considera-se que as aposições funcionam como atribuições de características, e não menções com relação de identidade. O tipo identidade é utilizado para anotar a correferência de identidade entre menções pronominais e nominais de referentes específicos, excluindo menções de entidades genéricas, subespecíficas ou abstratas (como classes/tipos de entidades ou eventos[37], ou hipóteses). O tipo apositivo, por sua vez, lida com relações apositivas, como a que ocorre em "Marcos, o presidente da empresa", em que “Marcos” seria anotado como tendo relação apositiva com “o presidente da empresa”. No *OntoNotes*, quando uma entidade referida por uma aposição é mencionada em outro ponto do texto, ao invés de adicionar a menção e sua aposição separadamente à corrente de correferência, a sequência contendo a construção apositiva inteira (menção + aposição) é incluída nessa corrente (dessa forma, criam-se correntes diferentes para a relação de aposição e a relação de correferência). Ou seja, usando o exemplo anterior, caso houvesse outra menção de “Marcos” no texto, essa teria relações de identidade com a menção “Marcos, o presidente da empresa”, e não com “Marcos” e “o presidente da empresa” separadamente. (BBN Technologies, 2024)

Para a anotação de predicados verbais, é anotado apenas o núcleo do sintagma verbal. Estes núcleos podem ser associados a outro verbo ou a nominalizações morfologicamente

---

[36] Acessadas em: https://data.mendeley.com/datasets/zmycy7t9h9/2/files/722501a7-2c8e-435d-9920-a0ebe2081f2a

[37] Como “buracos negros” em “buracos negros são objetos astronômicos”, ou “a morte de um familiar” em “a morte de um familiar é sempre um evento triste”, que não se referem a um buraco negro ou à morte de um familiar específico, mas a qualquer objeto ou acontecimento desse tipo.

relacionadas e sintagmas nominais que se refiram ao mesmo evento. Menções genéricas ou inespecíficas não são associadas a nenhum tipo de menção, a não ser pronomes dos quais sejam antecedentes (ou seja, são marcáveis como parte de uma relação anafórica, mas não parte de relações de correferência). Além disso, gentílicos não são considerados correferentes (ou seja, “Brasil” e “brasileiro” não são anotados como correferentes), assim como subpredicações (a relação entre “Eli” e “uma grande amiga” em “Jorge considera Eli uma grande amiga”). (BBN Technologies, 2024)

Expressões temporais podem ser anotadas como correferentes, incluindo expressões dêiticas de tempo, quando possível deduzir o período a que se referem. No entanto, partes de expressões temporais compostas por múltiplas datas, como "novembro" em "18 de novembro de 1987", não são anotadas isoladamente, sendo registradas apenas as expressões completas, de forma que o trecho “novembro” dentro desta menção não poderia ser associado a uma menção isolada de “novembro”. Não são anotados os fenômenos de anáfora indireta, ou com antecedentes coordenados. (BBN Technologies, 2024)

Foram feitas as seguintes escolhas e adaptações às diretrizes da OntoNotes:

a) Não serão anotadas relações de correferência entre eventos, apenas entre entidades.
b) A anotação será restrita a relações de correferência, abrangendo apenas menções referentes a entidades específicas e não incluindo relações unicamente anafóricas entre menções que não sejam correferentes. Ou seja, não serão anotadas expressões genéricas e nem pronomes que se refiram a elas.
c) As relações apositivas serão anotadas utilizando a etiqueta de tipo *appos*, proposta pelo formato CorefUD. Diferentemente do que ocorre no OntoNotes, contudo, no formato CorefUD tais relações são consideradas parte da corrente de correferência da entidade a que o aposto se refere, não havendo distinção prática entre relações de aposto e de identidade no momento da avaliação das anotações de um modelo. Dessa forma, não se faz uma segunda anotação da menção, contendo o elemento e seu aposto juntos, para sua inclusão na corrente de correferência, como é feito no OntoNotes. O elemento e seu aposto são anotados somente uma vez, separadamente, como pertencentes à mesma corrente de correferência.
d) Será feita a anotação de estruturas predicativas - que, no formato CorefUD, são marcadas utilizando o tipo *pred* - exceto nos casos em que as relações sejam baseadas em opiniões ou hipóteses. Apesar de o limite entre opinião e fato em um discurso ser turvo, especialmente em casos envolvendo atribuição de características a entidades,

> que podem ser implicitamente uma constatação subjetiva, e não factual, por parte do locutor, para o presente trabalho convencionou-se considerar como opiniões as relações marcadas por palavras de opinião, como "acha" e "considera", em frases como "Jorge acha João um sujeito estranho" e hipóteses as relações marcadas por verbos modais ou advérbios como "possivelmente" ou "provavelmente", como "João pode ser o culpado do crime".

O corpus revisado, de acordo com as diretrizes estabelecidas, resultou em 55 cadeias de correferência, com um total de 195 menções, em comparação com as 50 cadeias e 157 menções presentes no mesmo segmento do corpus original. Apesar de não ter sido uma mudança numericamente muito grande, a reorganização das menções pré-existentes e a inserção de novas menções, conforme elaborado na análise dos resultados descritos Seção 4, parece ter tido um impacto positivo na qualidade do corpus, tornando-o possivelmente mais adequado para uma avaliação mais fiel do desempenho de modelos de RC no português.

3.2 ELABORAÇÃO DOS *PROMPTS*

Estudos anteriores que exploram a resolução de correferência utilizando engenharia de *prompt* em LLMs (Sanh *et al.*, 2022 apud Yang *et al.*, 2022; Yang *et al.*, 2022; Gan *et al.*, 2024) não abordam a tarefa completa de RC (detecção, classificação e agrupamento de menções em cadeias de correferência). Salvo o modelo proposto por Le *et al*. (2023), boa parte dos modelos (Sanh *et al*., 2022 apud Yang *et al*., 2022; Perez *et al*., 2021 apud Anikina *et al.*, 2023; Lin *et al*., 2023; Yang *et al*., 2022; Gan *et al*. 2024) emprega apenas perguntas binárias com pares de menções previamente fornecidos, esquemas no estilo WSC ou estratégias similares, que possuem um formato diferente das tarefas tradicionais de resolução de correferência[38].

Este estudo visa avaliar o desempenho dos modelos testados na execução completa da tarefa de RC, contemplando todas as subtarefas envolvidas. O problema de RC foi abordado da seguinte forma: o *input* é o texto a ser analisado, e o *output* consiste em uma lista que apresenta as entidades identificadas no texto junto de suas respectivas menções, formatada de acordo com um padrão predefinido[39].

[38] Ver p.47.
[39] Descrito na p.85.

Foram inicialmente testadas diferentes opções de *prompts*, variando em complexidade, nível de detalhamento e forma de apresentação das tarefas. Esses testes preliminares foram realizados de maneira manual em um pequeno trecho de texto, com o objetivo de otimizar a identificação das abordagens mais eficazes. Essa estratégia permitiu ajustes rápidos nas instruções, ao dispensar o processamento de cada variação nos corpora de teste completos, reduzindo assim o tempo e os recursos necessários. Ao término desse processo, as duas combinações com melhor desempenho foram aplicadas aos corpora de teste escolhidos para a avaliação final dos modelos.

A seguir, detalha-se o processo de testagem e adaptação dos *prompts*, apresentando as variações testadas e algumas das respostas obtidas, que ilustram os principais tipos de erros identificados. Além disso, descrevem-se as estratégias empregadas para mitigar esses erros.

O segmento de texto utilizado para os testes preliminares foi:

> O apresentador Silvio Santos morreu neste sábado, aos 93 anos, e um dos clubes que prestaram homenagem nas redes sociais ao ícone da televisão brasileira foi o Flamengo. Curiosamente, o Rubro-Negro foi o último campeão do torneio criado e promovido pelo dono do SBT.
>
> Em 1995, em parceria com a empresa "Sport Promotion", o SBT inventou a Copa dos Campeões Mundiais, que reunia os times brasileiros que já haviam ganhado o Mundial Interclubes até então: Flamengo (1981), Grêmio (1983), São Paulo (1992 e 1993) e Santos (1962 e 1963). O torneio, em caráter amistoso, era disputado em julho, depois dos estaduais e antes de começar o Campeonato Brasileiro, e em duas sedes fixas.
>
> E valia tudo para promover a disputa. Por exemplo, Renato Gaúcho, que era do Fluminense e havia sido algoz do Flamengo na final do Carioca de 1995, jogou o torneio pelo Grêmio, onde tinha conquistado o título mundial em 1983, e enfrentou novamente o Flamengo na estreia. O Jornal do Brasil da época disse que o SBT também convidou Zico (aos 42 anos) e Raí (que estava no PSG) para reforçarem naquela edição o Rubro-Negro e o São Paulo, respectivamente, mas eles não jogaram.
>
> O São Paulo ganhou a competição em cima do Santos, e o SBT teve a média de 25 pontos de audiência com a sua Copa. Segundo contou o diretor de esportes do canal, Osmar de Oliveira, ao "Jornal do Brasil" na época, Silvio Santos estava satisfeito com o investimento no esporte e a ideia para o ano seguinte era convidar um time europeu. Mas em 1996 o torneio passou a entrar no calendário da CBF.[40]

[40] Texto disponível em: https://ge.globo.com/futebol/times/flamengo/noticia/2024/08/17/voce-sabia-flamengo-foi-campeao-da-ultima-edicao-do-torneio-criado-por-silvio-santos.ghtml

Inicialmente foram testados *prompts* simples e diretos, contendo uma breve descrição da tarefa e o texto a ser analisado, como “Liste todos os trechos do texto a seguir que se referem a uma mesma entidade”.

O uso de *prompts* com uma descrição muito ampla/genérica da tarefa a ser cumprida dá abertura a diferentes interpretações de como os detalhes dessa tarefa devem ser executados. Alguns dos erros mais frequentes nas respostas a essa abordagem foram[41]:

a) Inclusão de relações de anáfora indireta: os modelos frequentemente incluíam na corrente de correferência de uma entidade menções a uma parte dela, como ocorre no exemplo i, ou a um grupo ao qual fazem parte, como ocorre no exemplo ii.
   i) Resposta do modelo:
   SBT: “o dono do SBT”, “o SBT”, “o SBT”.
   ii) Resposta do modelo:
   O time Flamengo: “clubes que prestaram homenagem nas redes sociais ao ícone da televisão brasileira”, “o Flamengo”, “Flamengo”, “o Flamengo”.

b) Inclusão de antecedentes coordenados: frequentemente, os modelos inseriam na corrente de correferência de uma entidade uma expressão com mais de um antecedente, sendo essa entidade um deles, como no exemplo i, em que “eles” se refere aos jogadores Zico e Raí, simultaneamente.
   i) Resposta do modelo:
   Zico: “Zico”, “eles”.

c) Inclusão de anáfora zero: em alguns casos, os modelos incluíam anáforas zero na corrente de uma entidade, como no exemplo i.
   i) Resposta do modelo:

[41] Nos exemplos, extraídos das respostas dos modelos obtidas durante os testes preliminares, frequentemente há outros erros além daqueles que estão sendo analisados. Estes não são mencionados na análise, visando manter sua objetividade e simplicidade.

Renato Gaúcho: "Renato Gaúcho", "[ele] tinha conquistado", "[ele] enfrentou novamente o Flamengo".

d) Inclusão de conceitos genéricos: havia uma tendência dos modelos em associar conceitos genéricos ou categorias mais amplas à corrente de uma entidade específica, como a classe ou categoria a que pertence, como visto no exemplo i, em que "um time europeu" não se refere a uma entidade específica, mas a um elemento não-especificado, hipotético, de uma classe de entidades (a classe de todos os times europeus).

   i) Resposta do modelo:

   Time a ser convidado para o torneio: "um time europeu".

Embora não seja necessariamente indicativo de uma fraqueza na compreensão geral do texto por parte do modelo, e sim indicativo de uma compreensão mais ampla/holística, a inclusão das relações de anáfora indireta, antecedentes coordenados e anáfora zero nas respostas traz dificuldades técnicas para o processo de testagem proposto, dado que no formato CorefUD (bem como no formato do *Universal Anaphora*, o único outro formato que permite anotação de anáforas indiretas e antecedentes coordenados), essas relações devem ser anotadas de maneira distinta à de relação de correferência[42], o que dificulta consideravelmente a diferenciação e conversão desse tipo de resposta do modelo para esse formato. A inclusão de relações envolvendo expressões genéricas, por sua vez, também decorre de uma tentativa do modelo de captar as relações entre os elementos do discurso de forma mais abrangente, mas sua inclusão não caracteriza correferência ou anáfora e prejudica a avaliação do modelo pelos métodos utilizados, uma vez que acaba sendo anotada da mesma forma que as relações de correferência.

Dadas as limitações técnicas das ferramentas disponíveis para uma avaliação que considere todas essas relações, neste trabalho opta-se por abrir mão de uma análise mais completa em prol de uma análise mais confiável, de forma que o objetivo do trabalho é avaliar estritamente a capacidade de RC dos modelos. Assim, os *prompts* foram adaptados para evitar tais comportamentos indesejados. Para isso, foram incluídas instruções que especificam claramente quais relações devem ou não ser incluídas nas respostas, delimitando mais detalhada e especificamente o escopo da tarefa a ser executada.

[42] Formato de anotação descrito na p. 63.

Para isso, foi adicionada uma seção introdutória ao *prompt*, contendo uma descrição breve de fundamentos teóricos relacionados à RC – como os de modelo de discurso, entidade, referência, menção, anáfora, anáfora indireta, antecedentes coordenados e correferência, bem como uma explicação sobre o nível de especificidade que um conceito deve ter para que seja considerado uma entidade –, e instruindo à não-inclusão às respostas de conceitos genéricos e relações de anáfora indireta e antecedentes coordenados, orientando o modelo a focar exclusivamente nas relações de correferência.

Essa definição mais detalhada dos fundamentos e da tarefa contribuiu para uma redução significativa na frequência dos comportamentos indesejados descritos. No entanto, essa abordagem também resultou em um aumento na omissão de menções, deixando as listas de correferência consideravelmente menos completas – possivelmente pela divisão do "foco" do modelo entre a tarefa de identificação de menções e a de compreensão e aplicação dos conceitos descritos.

Além disso, optou-se por especificar um formato de resposta que permitisse a automatização do processo de conversão das respostas do modelo para o formato CorefUD. Esse processo é descrito na seção 3.4 e envolve a geração da resposta em uma lista estruturada e o uso de uma versão numerada do texto analisado. Para garantir essa formatação, foi adicionada uma seção ao final do *prompt* descrevendo o formato de resposta desejado. Dessa forma, o prompt passou a ter a seguinte estrutura.

Conforme observado por Gan *et al.* (2024) e Chiang *et al.* (2023, apud Gan *et al.*, 2024), a restrição no formato pode resultar em uma menor acurácia dos resultados quando se utilizando *prompts* em LLMs para tarefas de PLN. Essa redução na qualidade das respostas descrita pelos autores foi, de fato, observada ao se restringir o formato de resposta nos testes do presente trabalho, resultando em respostas ainda menos abrangentes.

Visando mitigar a redução na abrangência das respostas resultante tanto da restrição do escopo da tarefa quanto do formato de resposta, foi utilizada a estratégia de *Chain-of-Thought* (Wei *et al.*, 2023 *apud* Gan *et al.*, 2024), que permite que modelos realizem raciocínios complexos ao dividir o processo em etapas intermediárias de raciocínio (Prompt Engineering, 2024).

Dessa forma, a tarefa de listagem de menções foi estruturada em etapas sequenciais: uma etapa de identificação das unidades de interesse, uma etapa de reconhecimento de entidades e uma etapa de agrupamento de menções que se referem à mesma entidade, seguidas

por uma etapa de revisão do texto analisado, com o objetivo de localizar menções omitidas das entidades já identificadas.

Embora essa abordagem tenha melhorado a abrangência das respostas, o modelo voltou a realizar agrupamentos incorretos, como os observados nos exemplos dos itens *a*, *b* e *c*, mesmo com a definição já existente no *prompt* de que tipo de relações deveria ou não ser incluído na resposta. Para lidar com essa questão, foi adicionada uma etapa de revisão do agrupamento de menções na lista gerada.

Além disso, outro problema recorrente foi observado: em alguns casos, o modelo identificava uma menção com várias ocorrências idênticas e, à medida que avançava pelas etapas, esquecia que essas menções eram distintas, tratando-as como uma única instância. Um exemplo disso é o caso descrito no exemplo *e*, em que múltiplas menções de "o Flamengo" acabaram sendo ignoradas ao longo do processo.

e) Resposta do modelo:

[Na etapa de revisão do texto]

O time Flamengo:

1. *“o Flamengo”*
2. *“o Rubro-Negro”*
3. *“Flamengo”*
4. *“o Flamengo”*
5. *“o Flamengo”*

[Após a etapa de revisão do agrupamento]

O time Flamengo:

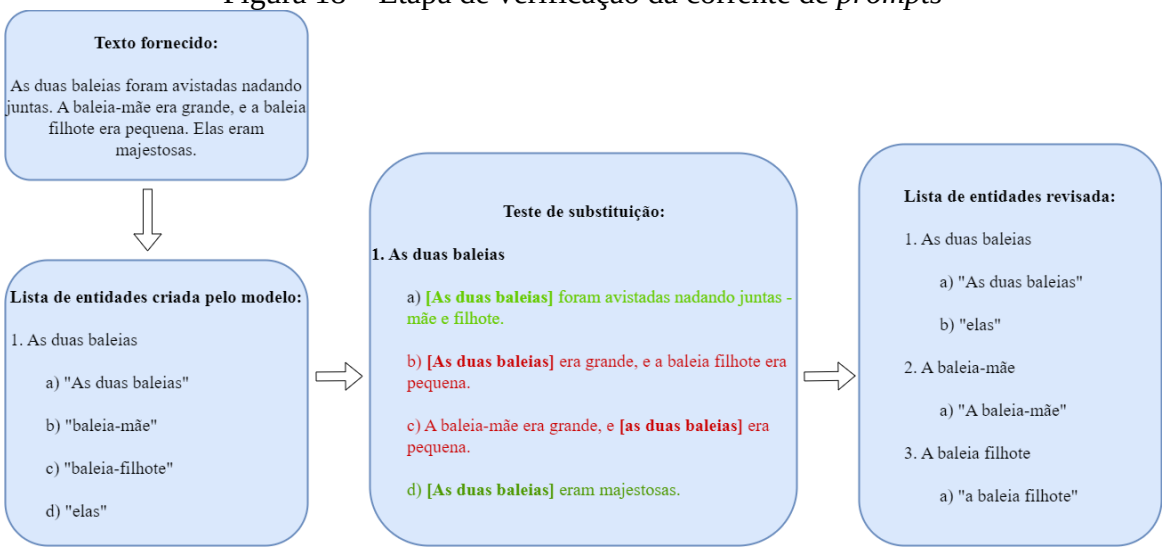
1. *“o Flamengo”*
2. *“o Rubro-Negro”*
3. *“Flamengo”*

Para mitigar essa questão, uma estratégia eficaz foi instruir que cada menção, em qualquer etapa do prompt, fosse sempre acompanhada do trecho exato em que aparece no texto.

A estratégia de inclusão dos trechos em que cada menção ocorre teve bom desempenho em evitar a perda de menções listadas – não houve mais esse tipo de ocorrência após sua inclusão –, mas a etapa de revisão não teve resultados tão satisfatórios: o modelo corrigia alguns erros de agrupamento, mas muitos eram ignorados.

Diante disso, foi elaborada uma variante dessa etapa envolvendo um teste de substituição. Dessa forma, como ilustrado na Figura 18, para cada menção associada a uma entidade, o modelo deve substituí-la no trecho em que ocorre pelo título dessa entidade na lista respondida e determinar se o sentido original da frase se mantém. Caso não se mantenha, a menção deve ser realocada para uma nova corrente.

Figura 18 – Etapa de verificação da corrente de *prompts*



Fonte: autora.

Essa modificação melhorou significativamente o agrupamento, mantendo uma boa abrangência nas respostas. No entanto, surgiram dois novos problemas: o modelo passou a cometer erros de formatação – como, por exemplo, os erro observado no exemplo *f*, em que cada palavra de uma menção numerada separadamente das demais, e *g*, em que o modelo numera todo o trecho em que a menção ocorre, e não apenas as palavras da menção em si - e, ocasionalmente, a alucinar ou inventar menções que não estavam presentes no texto, conforme ocorre no exemplo *h*, em que o modelo inventa a menção “o torneio” no trecho “a ideia para o ano seguinte era convidar um time europeu [para o torneio]”.

f) Resposta do modelo:

Silvio Santos:

1. O [1] apresentador [2] Silvio [3] Santos [4], em “O apresentador Silvio Santos morreu neste sábado”.

g) Resposta do modelo:

Silvio Santos:

1. “O apresentador Silvio Santos”, em “O apresentador Silvio Santos morreu neste sábado” [1,2,3,4,5,6,7].

h) Resposta do modelo:

O campeonato organizado pelo SBT:

1. “o torneio”, em “o último campeão do torneio criado e promovido pelo dono do SBT”

2. “a Copa dos Campeões Mundiais”, em “o SBT inventou a Copa dos Campeões Mundiais”

3. “O torneio”, em “O torneio, em caráter amistoso, era disputado em julho”

4. “aquela edição”, em “o SBT também convidou Zico (aos 42 anos) e Raí (que estava no PSG) para reforçarem naquela edição o Rubro-Negro e o São Paulo”

5. “a competição”, em “O São Paulo ganhou a competição em cima do Santos”

6. “o torneio”, em “a ideia para o ano seguinte era convidar um time europeu [para o torneio]”

7. “o torneio”, em “em 1996 o torneio passou a entrar no calendário da CBF”

A quantidade excessiva de instruções fazia com que o modelo falhasse em considerar/memorizar simultaneamente todos os aspectos mencionados, resultando em uma situação em que a melhoria em um elemento comprometia o desempenho nos demais, forçando uma escolha entre abrangência e confiabilidade nas respostas. Para resolver essa questão, foi adotada a estratégia de *prompt chaining*. Essa técnica consiste em dividir uma tarefa complexa em subtarefas, submetendo cada uma delas como um *prompt* a um modelo. O *output* de cada *prompt* é, então, utilizado como *input* para o próximo, criando uma cadeia de operações. Isso permite reduzir a quantidade de informações e instruções que o modelo precisa processar de uma só vez, melhorando a confiabilidade e facilitando a identificação e correção de problemas nas respostas (Prompt Engineering, 2024) Dessa forma, as etapas de identificação de unidades de interesse, reconhecimento de entidades, agrupamento de menções, revisão do texto, revisão de agrupamento e formatação da resposta foram submetidas ao modelo em prompts separados,

assegurando que a execução de uma tarefa não fosse prejudicada pela execução simultânea de outra.

Essa abordagem apresentou bons resultados, alcançando um equilíbrio satisfatório entre confiabilidade e abrangência. Se mostrou uma solução interessante especialmente para a questão da restrição de formato descrita por Gan *et al.* (2024) e Chiang *et al.* (2023, apud Gan *et al.*, 2024), uma vez que a etapa de formatação é realizada apenas após a conclusão de todas as demais etapas da tarefa, não sendo influenciada por elas.

Com base nesses testes, nos problemas observados e nas soluções implementadas, foram finalmente testadas duas sequências de *prompts* — uma mais simples e outra mais complexa — que serão denominadas Abordagem A e Abordagem B, descritas nas seções seguintes.

3.2.1 Abordagem A

Nesta abordagem, a tarefa é realizada de forma relativamente direta. O modelo recebe o texto e retorna uma lista de entidades e suas menções, sem etapas intermediárias. Após essa fase, são realizadas duas etapas de revisão: a etapa de busca no texto por menções ignoradas e a etapa de revisão do agrupamento das menções (sua versão simplificada, sem o uso do teste de substituição). Por fim, é feita a conversão da lista para o formato desejado. Foram avaliadas versões *0-shot* e *few-shot* dessa abordagem. A abordagem segue três etapas principais, contando com um total de 4 *prompts*:

1) Formação de uma lista de entidades e menções com base no texto analisado (*prompt 1*).
2) Revisão da lista elaborada.
    i) Busca por menções ignoradas (*prompt 2*).
    ii) Revisão do agrupamento das menções (*prompt* 3).
3) Formatação da lista final (*prompt 4*).

3.2.2 Abordagem B

Nesta abordagem, a tarefa inicial de formação da lista é subdividida em mais etapas antes da fase de revisão. As etapas incluem a identificação de unidades de interesse, definição de referencialidade das menções (ou seja, se são elementos do texto que se referem a conceitos suficientemente específicos para que sejam considerados menções a uma entidade),

agrupamento das menções às entidades, revisão – dividida em uma etapa de busca no texto por menções ignoradas e uma etapa de revisão do agrupamento das menções com o uso do teste de substituição – e conversão final para o formato desejado. Ao todo, a execução da tarefa de RC foi organizada em seis *prompts*:

a) Formação da lista
    i) Identificação de unidades de interesse no texto (*prompt* 1).
    ii) Definição de referencialidade dos trechos identificados (*prompt* 2).
    iii) Identificação das entidades e agrupamento das menções (*prompt* 3).
b) Revisão da lista elaborada.
    i) Busca por menções ignoradas (*prompt* 4).
    ii) Revisão do agrupamento das menções por meio do teste de substituição (*prompt* 5).
c) Formatação da lista final (*prompt* 6).

A estrutura e descrição detalhadas de cada um dos *prompts* desenvolvidos para a abordagem 1 e 2 pode ser conferida no Apêndice A e B, respectivamente.

3.3 MODELOS

Para a resolução dos corpora de teste definidos, serão empregados o ChatGPT-4o e uma versão customizada do ChatGPT-4, adaptada para a tarefa de resolução de correferência conforme as diretrizes de anotação delimitadas na seção 2.1.3.3.

3.3.1 ChatGPT-4o

Lançado em maio de 2024, o ChatGPT-4o é o modelo mais recente da OpenAI, que traz avanços significativos em relação às versões anteriores. Ele oferece desempenho equivalente ao GPT-4 Turbo em tarefas de processamento de texto em inglês e geração de código, mas tem melhorias notáveis no processamento de textos em outros idiomas. No *benchmark* MMLU[43], o ChatGPT-4o alcançou uma precisão de 88,7%, superando os 86,5% do GPT-4 Turbo e os 86,4% do GPT-4. (OpenAI, 2024b).

[43] O MMLU (Massive Multitask Language Understanding) é um *benchmark* criado para avaliar a precisão de modelos de linguagem em uma ampla gama de tarefas que abrangem 57 disciplinas diferentes, como matemática,

3.3.2 GPT customizado

Em novembro de 2023, a OpenAI introduziu uma ferramenta de criação de versões customizadas do ChatGPT, denominadas *GPTs* (OpenAI, 2024c). Baseados no modelo ChatGPT-4, os GPTs permitem a combinação de diferentes componentes, denominados *GPT internals*, que incluem *system prompts*, arquivos de conhecimento e funcionalidades externas (Liu *et al.*, 2024; Zhang *et al.*, 2024).

Os ***system prompts*** são, de acordo com Xu *et al.* (2024) instruções que funcionam como um "prefixo padrão implicitamente adicionado ao *input* do usuário" durante a fase de inferência dos LLMs. Esses *prompts* operam de forma subjacente às interações com o usuário e contêm diretrizes detalhadas que ajudam o modelo a interpretar melhor o contexto e fornecer respostas mais adequadas às solicitações recebidas. Pape *et al.* (2024) apontam que o uso de *system prompts* com instruções detalhadas pode transformar LLMs genéricas em ferramentas específicas com uma demanda de recursos mínima e sem a necessidade de ajustes finos com aprendizado supervisionado.

Nos GPTs, o *system prompt* pode ser inserido na seção *Instruções*, que permite a especificação de instruções e anotações críticas, possibilitando que os GPTs personalizados interajam com fontes de informação com maior precisão (Liu *et al.*, 2024). Além da seção de instruções, a função *Conhecimento* permite ao usuário carregar arquivos contendo informações como diretrizes, suplementação teórica e conhecimento especializado em um determinado domínio (Liu *et al.*, 2024).

As funcionalidades externas, por sua vez, incluem navegação na *web*, geração de imagens e interpretação de código. A navegação na web permite que o GPT use o *Bing* para buscar conteúdo ou interaja com sites pré-definidos. A geração de imagens com DALL·E cria imagens a partir de texto, que podem ser exibidas e baixadas, e o interpretador de código executa programas em Python no sistema Linux. (Tao *et al.*, 2023)

Para a criação do GPT utilizado neste trabalho[44], foram utilizados os seguintes elementos:

a) Instruções/*system prompt*:

---

história, ciência da computação e direito. Ele foi projetado para testar o conhecimento geral e a habilidade de resolução de problemas dos modelos, exigindo uma combinação de conhecimento de mundo e raciocínio avançado. (Hendrycks, 2020)

[44] Disponível para acesso em: https://chatgpt.com/g/g-UPoLvr0Yl-especialista-em-resolucao-de-correferencia

i) Definição da tarefa: o modelo deve receber um texto e retornar uma lista de entidades e suas respectivas menções, seguindo as diretrizes fornecidas.

ii) Etapas do processo:

1) Identificação e listagem de entidades e suas menções com base no texto fornecido e nas instruções do usuário.
2) Revisão da lista elaborada.
3) Criação de versão numerada da lista de entidades e suas menções elaborada e revisada anteriormente, conforme formatação especificada.

b) Conhecimento:

i) Fundamentação teórica: documento contendo definição clara e concisa dos conceitos descritos na seção 2.1 e 2.2 deste trabalho (modelo de discurso, referência, anáfora, catáfora, correferência, resolução de correferência, etc.), que servem como base de conhecimento especializado.

ii) Diretrizes: documento contendo as diretrizes de anotação de correferência descritas na seção 2.1.3.3 deste trabalho.

iii) Formatação: breve descrição do formato esperado para o *output*[45].

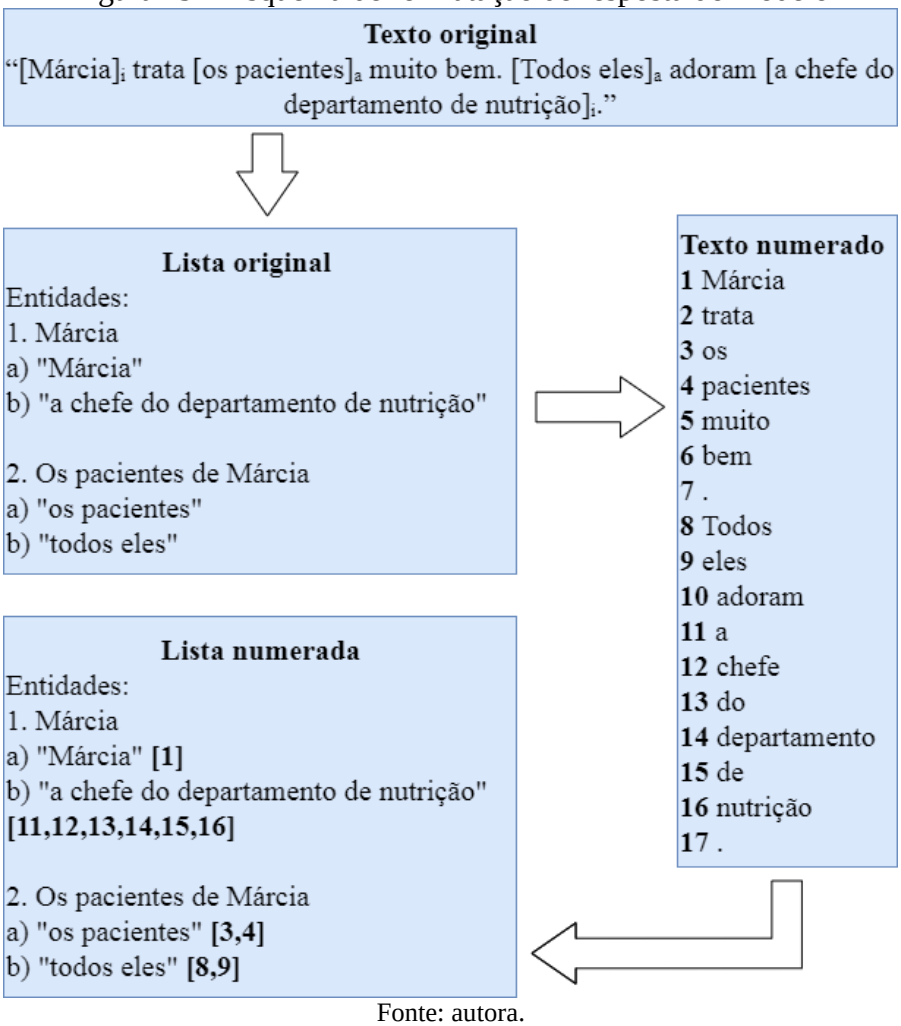
As funcionalidades externas (acesso à *web*, geração de imagens e interpretação de código) não foram utilizadas.

3.4 CONVERSÃO DAS RESPOSTAS PARA O FORMATO COREFUD

A conversão das respostas dos modelos para o formato CorefUD foi conduzida de forma semiautomática. A estratégia adotada envolveu a solicitação do *output* dos modelos no formato de uma lista, contendo as entidades e suas respectivas menções, cada uma acompanhada pela numeração correspondente de suas palavras entre colchetes conforme uma versão numerada do texto analisado. O processo completo é exemplificado na Figura 19.

[45] Descrito na seção 3.4.

Figura 19 – Esquema de formatação de resposta do modelo



Fonte: autora.

Em seguida, as informações numéricas foram extraídas e processadas, gerando as anotações correspondentes no formato CorefUD. Por fim, foi feita uma revisão manual para corrigir eventuais falhas de conversão, dado que alguns *outputs* continham imprecisões em sua formatação. A ferramenta *MoveHead*[46] foi, então, utilizada para atribuir núcleos às menções já convertidas para o formato CorefUD.

3.5 PONTUAÇÃO

Embora as métricas tradicionais de precisão, *recall* e F^1 é sejam amplamente utilizadas na avaliação de tarefas de classificação binária, como a detecção de menções (Yu *et al*., 2020; Peng *et al*., 2015 apud Liu *et al*., 2023) e na seleção binária para ligação de menções (Kocijan

[46] Descrita na p.65.

et al., 2019; Attree, 2019 apud Liu *et al*., 2023), elas apresentam limitações significativas quando aplicadas a modelos que operam com cadeias de correferência, que exigem uma visão mais global da interação entre menções (Liu *et al*., 2023). Como exposto na seção 2.6, essas métricas tradicionais são projetadas para tarefas de classificação binária, enquanto que o funcionamento de sistemas de RC envolvendo agrupamento de menções nem sempre tem um componente de classificação binária óbvio/diretamente avaliável por esse tipo de métrica.

Para a avaliação de índices análogos aos de precisão, *recall* e F^1 nesse tipo de sistemas, é necessário que se faça uma decisão sobre o quê, exatamente, será avaliado em termos de positivo, negativo, verdadeiro e falso. Algumas métricas avaliam as conexões entre menções, enquanto outras avaliam a associação das menções a cada grupo (Sukthanker *et al.*, 2020).

Diante disso, para a avaliação dos modelos testados neste trabalho, será empregada uma combinação de métricas amplamente utilizadas na área para esse tipo de tarefa – prática comum na área de RC (Liu *et al.*, 2023). Para isso, será utilizado o CorefUD *scorer*[47], ferramenta oficial das tarefas compartilhadas da CRAC para resolução de correferência multilíngue, que suporta anotações no formato CorefUD 1.0.

O *scorer* trabalha com correspondências exatas, parciais e de núcleo das menções, e suporta todas as principais métricas usadas para correferência, como MUC, B^3, CEAF, BLANC e LEA. O avaliador é limitado à avaliação da resolução correferências, não contemplando os casos de antecedentes coordenados, anáfora indireta e outras relações (GitHub 2024). Além disso, as métricas suportadas pelo *scorer* (sendo elas as mais comumente utilizadas na área) não calculam índices de acurácia. Dessa forma, embora se trate um índice útil para a análise de sistemas de PLN, o presente trabalho não contará com a análise da acurácia dos modelos testados, dada a inexistência de métodos amplamente utilizados para o cálculo desse índice na área de RC, de maneira que isto demandaria a elaboração de um método próprio para tal.

No presente trabalho, o método de correspondência entre as menções dos corpora de teste e as menções respondidas pelos modelos testados foi o de núcleos. Na correspondência de núcleos, duas menções são consideradas correspondentes se seus núcleos corresponderem a tokens idênticos, e os intervalos completos das menções são ignorados. No entanto, se houver mais de uma menção com o mesmo núcleo, o modelo aplica regras de desambiguação, uma vez que nenhuma menção do corpus original pode corresponder a mais de uma menção do modelo de resposta ou vice-versa. Nesta situação, os intervalos completos das menções podem ser

[47] Descrita na seção 2.6.7, p.60.

considerados para fins de desambiguação. Para obter uma única menção correspondente nos casos em que mais de uma menção *n* do modelo de resposta tenha o mesmo núcleo que uma menção *m* do corpus original, as seguintes regras de desambiguação são obedecidas:

a) Escolhe-se a menção que se sobrepõe a *m* com a menor diferença proporcional de intervalo.
b) Se ainda restar mais de uma menção *n*, escolhe-se a que começa mais cedo no documento.
c) Se ainda restar mais de uma menção *n*, escolhe-se a que termina mais cedo no documento. (GitHub 2024)

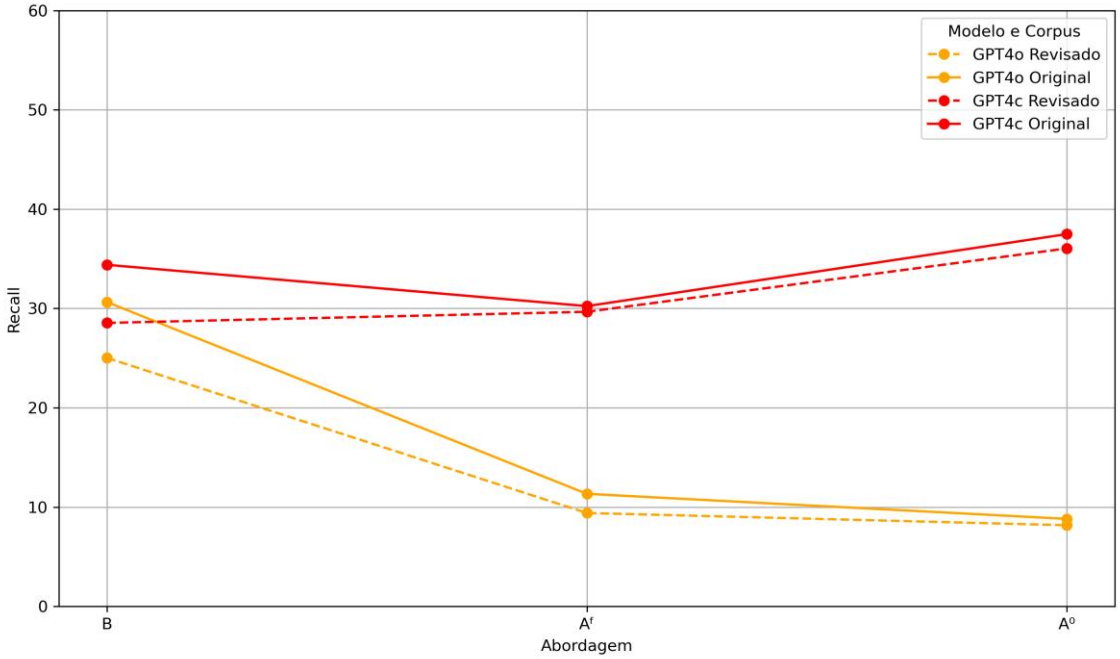
O método de correspondência de núcleos foi escolhido por ser mais permissivo, considerando as possíveis imprecisões da formatação das respostas dos GPTs e, portanto, a possibilidade de divergências entre o corpus original e os modelos testados na definição do *span* das menções mas que não representam uma falha em sua capacidade de compreender correferências.

4 RESULTADOS

As Figuras 20 e 21, contêm, respectivamente, gráficos comparativos das médias dos índices de *recall* e precisão de todas as métricas nas quais os modelos foram avaliados por meio do CorefUD *scorer*. São elas: B^3, $CEAF_e$, CEAFm, BLANC, LEA e MOR (resultados de cada métrica no Apêndice C).

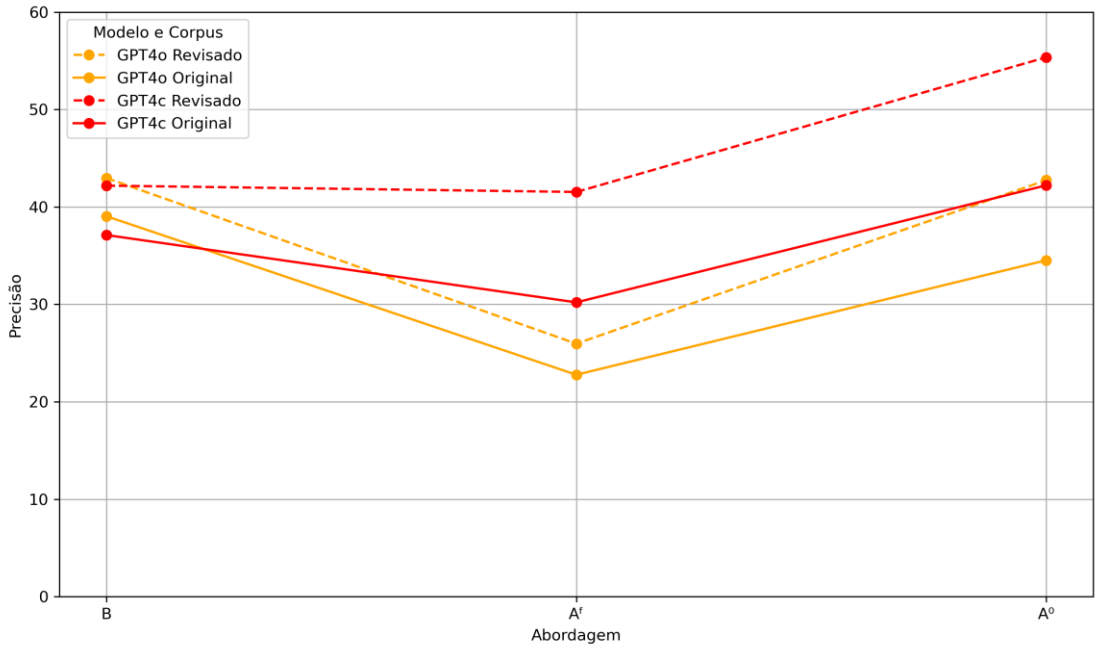
Em geral, o ChatGPT-4o apresentou desempenho semelhante nas abordagens B e A *zero-shot* em termos de precisão, e um desempenho pior na abordagem A *few-shot*. Nas abordagens A *few-shot* e A *zero-shot*, seu desempenho foi substancialmente inferior nos índices de *recall* quando comparado aos da abordagem B. Em termos de precisão, o GPT customizado teve seu pior desempenho na abordagem A *few-shot*, e seu melhor desempenho na abordagem A *zero-shot*. O mesmo ocorre nos índices de *recall*.

Figura 20 – Gráfico comparativo da média de *recall* dos modelos e abordagens testados



Legenda: GPT4c = GPT customizado. GPT4o = ChatGPT-4o.
B = Abordagem B. A_0 = Abordagem A *zero-shot*. A^f = Abordagem A *few-shot*.
Fonte: autora.

Figura 21 – Gráfico comparativo da média de precisão dos modelos e abordagens testados



Legenda: GPT4c = GPT customizado. GPT4o = ChatGPT-4o.
B = Abordagem B. A_0 = Abordagem A *zero-shot*. A^f = Abordagem A *few-shot*.
Fonte: autora.

Na abordagem B, o GPT customizado teve desempenho semelhante ao do ChatGPT-4o em termos de precisão e levemente melhor em termos de *recall*. Nas demais abordagens, teve desempenho consideravelmente melhor em ambas as métricas, especialmente nos índices de *recall*. Em geral, a combinação com os melhores resultados foi o uso do GPT customizado com a abordagem A *zero-shot* (A0), especialmente na resolução do corpus revisado. No caso do ChatGPT-4o, a abordagem com melhores resultados foi a B. A abordagem A *few-shot* não se mostrou vantajosa para nenhum dos modelos.

De modo geral, o uso de uma mesma combinação de modelo e abordagem obteve índices de *recall* semelhantes na resolução dos dois corpora – com exceção da abordagem B, na qual os modelos apresentaram resultados levemente melhores no corpus original –, como observa-se na Figura 20. Entretanto, houve um aumento significativo nos índices de precisão na resolução do corpus revisado em comparação à do original, especialmente no GPT customizado (Figura 21). Esse cenário pode ser um reflexo da natureza das revisões feitas no corpus Summ-it++, que, entre outras melhorias, adicionaram correntes e menções que haviam sido ignoradas no corpus original. Na resolução de um corpus anotado de maneira incompleta, as previsões corretas de um modelo podem ser injustamente penalizadas por conta da ausência dessas menções nas anotações de referência. Dessa forma, esses resultados podem indicar que

a revisão do Summ-it++ foi pertinente e eficaz, contribuindo de maneira relevante para um corpus mais completo.

A adição de exemplos à abordagem A resultou consistentemente em notas mais baixas, possivelmente porque os exemplos elaborados não foram eficazes. Ambos os modelos mostraram-se bastante sensíveis aos *prompts*, demonstrando variação considerável nos resultados conforme a abordagem utilizada. Dessa forma, o emprego de uma engenharia de *prompt* mais refinada nesses modelos pode alcançar resultados vantajosos, incluindo, entre outras coisas, a elaboração de exemplos mais adequados para as abordagens *few-shot*. Considerando o corpus revisado como uma ferramenta potencialmente mais precisa para avaliar o desempenho real dos modelos, observa-se uma discrepância entre as métricas de *recall* e precisão, com o *recall* apresentando valores consideravelmente mais baixos. Mesmo com o emprego de uma etapa de revisão de resgate de menções nas correntes de *prompts* desenvolvidas, os modelos parecem ter adotado uma abordagem mais conservadora. Diante disso, pode ser benéfica a implementação de estratégias que visem tornar os critérios dos modelos mais abrangente, a fim de otimizar a identificação de menções relevantes.

Tabela 17 – Comparação dos resultados obtidos pelos modelos deste trabalho aos dos resultados de outros modelos de RC do português

Modelo		Corpus	MUC			B^3			CEAFe			CoNLL
			P	R	F^1	P	R	F^1	P	R	F^1	F^1
Fonseca et al. (2017)		Summ-it++	42,30	**53,60**	47,30	38,70	**50,80**	43,90	45,6	52,8	48,9	46,70
Fonseca et al. (2018)	Sem informação semântica	Summ-it++	**58,80**	44,40	**50,60**	**59.3**	41.7	**49.0**	**53.7**	54.2	**54.0**	**51.2**
	Com informação semântica		45,10	52,10	48,30	43.8	49.0	46.5	45.7	**57.4**	50.9	48.6
Nosso modelo	GPT4o + B	Summ-it++	37,50	31,13	34,02	38,35	29,23	33,17	48,28	37,66	42,31	36.50
	$GPT4^c + A_0$		**42,86**	**39,62**	**41,18**	**40,78**	**36,84**	**38,71**	**49,63**	**43,67**	**46,46**	**42.11**
Nosso modelo	GPT4o + B	Summ-it++ revisado	38,64	24,29	29,82	42,28	22,06	28,99	49,50	36,43	41,97	33.60
	$GPT4^c + A_0$		56,12	39,29	46,22	53,94	33,75	41,52	54,47	45,22	49,41	45.72

Legenda: $GPT4^c$ = GPT customizado. A^0 = Abordagem A *zero-shot*.
B = Abordagem B.
Fonte: autora.

A Tabela 17 compara o desempenho observado no uso do GPT customizado e ChatGPT-4o com as abordagens A *zero-shot* e B, respectivamente – sendo essas combinações as que apresentaram os melhores resultados nos testes do presente trabalho – aos dos modelos de Fonseca *et al.* (2017) e Fonseca *et al.* (2018) na resolução do corpus Summ-it++. Enquanto

que a variação sem informação semântica do modelo de Fonseca *et al.* (2018), atual estado da arte na resolução do Summ-it++, alcançou uma pontuação de 51,2 na métrica CoNLL, o nosso modelo de melhor desempenho na resolução do corpus original, o GPT customizado com a abordagem A *zero-shot*, apresentou uma pontuação 17,75% inferior, de 42,11.

Tabela 18 – Comparação dos resultados do presente trabalho com os de Le *et al*. (2023)

Modelo		Corpus	*CoNLL Score*
Le *et al*. (2023)	ChatGPT-4	OntoNotes 5.0^{en}	88.4
	InstructGPT	AnCora^{es}	**42.2**
		AnCora^{ca}	**41.9**
Nosso modelo	GPT4o + B	Summ-it++	36.50
	GPT4^{c} + A^{0}		**42.11**
	GPT4o + B	Summ-it++ revisado	33.60
	GPT4^{c} + A^{0}		**45.72**

Legenda: GPT4^{c} = GPT customizado. A^{0} = Abordagem A *zero-shot*.
B = Abordagem B. ^{ca} = catalão. ^{es} = espanhol. ^{en} = inglês.
Fonte: autora.

A Tabela 18, por sua vez, apresenta uma comparação entre os resultados deste estudo e os de Le *et al*. (2023) em experimentos semelhantes. Diferentemente de outros estudos que exploram o uso de *prompting* para a RC, Le *et al*. (2023) adota uma abordagem mais próxima da tarefa tradicional de RC. Assim como nos testes realizados neste trabalho, os autores testam os modelos na anotação de correferência em textos completos, em vez de empregar apenas perguntas binárias com pares de menções previamente fornecidos, esquemas no estilo WSC ou estratégias similares (Perez *et al*., 2021 apud Anikina *et al.*, 2023; Lin *et al*., 2023; Yang *et al*., 2022; Gan *et al*. 2024). Além disso, conforme apontado por Recasens *et al*. (2010), os corpora utilizados por Le *et al*. (2023) seguem diretrizes de anotação de correferência semelhantes às do OntoNotes, que serviram de base para a revisão do Summ-It++, além de serem em línguas próximas ao português: o espanhol e o catalão.

Embora uma comparação direta não seja viável devido às diferenças nas abordagens, corpus e línguas entre os dois trabalhos, as similaridades observadas podem sugerir uma relação interessante. O estudo de Le *et al.* (2023) evidenciou que uma identificação precisa de menções exerce um impacto significativo nos resultados finais do uso de *prompting* para a resolução de correferência. Considerando as semelhanças do sistema proposto por Le *et al*. (2023) aos experimentos realizados neste estudo, e tendo em vista que obtiveram resultados semelhantes

utilizando versões mais antigas do ChatGPT – resultados que não superam os de atuais modelos envolvendo aprendizado supervisionado, mas que superam os resultados de outros modelos não supervisionados –, pode ser interessante explorar a integração do fornecimento prévio de identificação de menções às abordagens testadas neste estudo, a fim de investigar possíveis impactos positivos em seu desempenho.

Por último, visando complementar a análise dos resultados com uma perspectiva linguística, utilizou-se o módulo *corefud.Stats* do Udapi[48], que calcula o número total de entidades e menções presentes em um corpus, além de fornecer informações sobre o tamanho médio das menções e a distribuição das categorias gramaticais entre os seus núcleos.

Por meio dessa ferramenta, buscou-se investigar possíveis interações entre a frequência de diferentes classes gramaticais nos núcleos das menções anotadas e os índices médios de *recall* e precisão obtidos na resolução dos diferentes corpora. As informações sobre a distribuição da classe gramatical dos núcleos das menções na resolução dos corpora original e revisado por cada combinação de modelo e abordagem podem ser observadas nas Tabelas 19 e 20, respectivamente. No entanto, não foi identificada nenhuma correlação clara, linear ou monotônica entre essas variáveis e os índices de *recall* e precisão. É importante destacar que, devido ao tamanho reduzido da amostra analisada, é difícil determinar se essa ausência de correlação é confiável ou significativa.

Tabela 19 - Distribuição da classe gramatical dos núcleos de menções na resolução do corpus original pelos modelos testados

Modelo		CoNLL Score	Precisão	Recall	Substantivo	Pronome	Nome próprio	Determinante	Adjetivo	Verbo	Advérbio	Numeral	Outros
GPT4o	B	36,5	39,04	30,64	64,92%	7,03%	23,38%	1,83%	0%	0,58%	0%	4,47%	0%
	Af	14,09	22,77	11,33	32,08%	1,49%	24,63%	16,40%	4,46%	8,21%	0%	3,73%	8,94%
	A^0	15,34	34,52	8,81	33,69%	0,98%	39,59%	14,85%	4,94%	0,98%	0%	0%	4,94%
GPT4^c	B	37,41	37,11	34,39	20,98%	8,32%	8,78%	47,33%	2,44%	0,97%	0,48%	3,90%	6,83%
	Af	32,03	30,2	30,23	39,27%	7,33%	22%	17,81%	3,67%	0,52%	0%	5,75%	3,67%
	A^0	42,11	42,21	37,48	59,79%	4,27%	30,25%	2,40%	0%	0,65%	0%	2,60%	0%

Fonte: autora.

[48] Framework em Python para o processamento de dados das Universal Dependencies. Disponível em: https://github.com/udapi/udapi-python/tree/master.

Tabela 20 – Distribuição da classe gramatical dos núcleos de menções na resolução do corpus revisado pelos modelos testados

Modelo		CoNLL Score	Precisão	Recall	Substantivo	Pronome	Nome próprio	Determinante	Adjetivo	Verbo	Advérbio	Numeral	Outros
GPT4o	B	33,6	42,96	25,02	64,92%	7,03%	23,38%	1,83%	0%	0,58%	0%	4,47%	0%
	Af	12,42	25,96	9,4	32,08%	1,49%	24,63%	16,40%	4,46%	8,21%	0%	3,73%	8,94%
	A^0	14,17	42,74	8,17	33,69%	0,98%	39,59%	14,85%	4,94%	0,98%	0%	0%	4,94%
GPT4c	B	34,38	42,18	28,54	20,98%	8,32%	8,78%	47,33%	2,44%	0,97%	0,48%	3,90%	6,83%
	Af	34,59	41,53	29,66	39,27%	7,33%	22%	17,81%	3,67%	0,52%	0%	5,75%	3,67%
	A^0	45,72	55,35	36,04	59,79%	4,27%	30,25%	2,40%	0%	0,65%	0%	2,60%	0%

Fonte: autora.

Além disso, uma análise detalhada da frequência de núcleos de diferentes classes gramaticais entre os erros cometidos pelos modelos poderia contribuir para identificar se os modelos apresentam maior propensão a falhas na resolução de correferências envolvendo menções com determinados tipos de núcleo. Contudo, não foi encontrada nenhuma ferramenta disponível que permitisse a realização dessa análise de forma automatizada.

5 CONCLUSÃO

O presente trabalho avaliou o uso de engenharia de *prompt* para a tarefa de resolução de correferência de entidade na língua portuguesa em dois grandes modelos de linguagem pré-treinados: o ChatGPT-4o e uma versão do ChatGPT-4 customizada para a resolução de correferência. Essa customização foi feita por meio da ferramenta de customização disponibilizada pela OpenAI, utilizando um *system prompt* e arquivos de conhecimento baseados em fundamentos teóricos sobre correferência e nas diretrizes de anotação desejadas.

A elaboração dos *prompts* empregados nos testes empregou estratégias simples de engenharia de prompt e a técnica de *prompt chaining* - resultando na criação de duas séries de *prompts few* e *zero-shot*.

Foram utilizados dois corpora de teste: um segmento equivalente a 10% do corpus Summ-it++, o segundo maior corpus com anotação de correferência em português, e uma versão revisada desse mesmo segmento. A revisão consistiu em ajustar as anotações de correferência existentes, conforme diretrizes baseadas nas do OntoNotes, com o objetivo de corrigir erros e aumentar a consistência das anotações originais. Ambos os conjuntos também foram harmonizados às diretrizes de formatação da iniciativa CorefUD, garantindo acesso a uma gama mais ampla e atualizada de ferramentas da área.

Até onde se tem conhecimento, este trabalho representa a primeira aplicação da engenharia de *prompt* na tarefa de resolução de correferência na língua portuguesa. Além disso, introduziu o uso dessa estratégia na tarefa completa de resolução de correferência, incluindo todas as suas subtarefas, sem o fornecimento prévio de menções identificadas aos modelos testados.

Embora os resultados obtidos não tenham superado o estado da arte na resolução do corpus Summ-it++ em sua versão completa — representado pelo modelo de Fonseca et al. (2018) —, o desempenho alcançado foi relevante. A abordagem proposta atingiu um CoNLL score apenas 17,75% inferior ao modelo de referência, o que sugere um resultado promissor, especialmente considerando a simplicidade do modelo utilizado.

Além disso, a estratégia de *prompt chaining* empregada na criação das sequências de prompts testadas mostrou-se eficaz para mitigar a redução da qualidade das respostas em grandes modelos de linguagem diante de restrições no formato das respostas, conforme descrito por Gan *et al.* (2024) e Chiang *et al*. (2023, *apud* Gan *et al*., 2024). Ao dividir a resolução da tarefa em si e a etapa de formatação em dois prompts distintos, essa estratégia minimizou a influência mútua entre essas etapas, resultando em uma execução mais eficiente e consistente.

A customização do ChatGPT-4 também apresentou resultados promissores. O modelo customizado alcançou um CoNLL score máximo cerca de 15% superior ao score máximo alcançado pelo ChatGPT-4o no corpus original e cerca de 36% superior no corpus revisado. Esses números são especialmente relevantes considerando que o ChatGPT-4o é um modelo mais recente e avançado. Isso demonstra que o treinamento extremamente simples aplicado ao modelo customizado foi eficaz e trouxe ganhos consideráveis de desempenho, indicando o potencial dessa abordagem em aplicações futuras.

Adicionalmente, os índices de precisão consideravelmente mais altos observados na resolução do corpus revisado em relação à do original, sem prejuízo nos valores de *recall*, sugerem que o corpus revisado é possivelmente mais completo e consistente em relação ao original. Esse resultado reforça a pertinência do processo de revisão realizado, demonstrando que os ajustes nas anotações de correferência possivelmente contribuíram para a melhoria da qualidade e da usabilidade do corpus em tarefas de resolução de correferência.

O presente trabalho apresenta limitações que abrem espaço para aprimoramentos em estudos futuros. Entre elas, destaca-se a não inclusão da resolução de anáforas indiretas ou com antecedentes coordenados, sendo abordadas apenas anáforas de antecedente único que coocorrem com correferência. Essa escolha foi motivada pela complexidade envolvida na anotação e avaliação desses fenômenos no formato utilizado. No entanto, esses fenômenos são fundamentais para a compreensão de um discurso, e sua inclusão em testes futuros poderia expandir e aprofundar os resultados obtidos. Além disso, o modelo também não foi testado para a resolução de correferência de eventos, devido à reconhecida complexidade dessa tarefa. Investigar sua aplicação nesse contexto pode ser, também, uma direção relevante e promissora.

Outro aspecto a ser considerado é o número limitado de modelos testados, restrito a apenas dois modelos da mesma família, devido às limitações de recursos e tempo disponíveis. Inicialmente, pretendia-se incluir LLMs voltados ao português, como os da família Sabiá, mas testes preliminares mostraram que esses modelos não conseguiam produzir respostas no formato requerido. Dessa forma, além da expansão dos testes para outros modelos, o desenvolvimento de uma estratégia mais eficiente para a conversão das respostas dos LLMs ao formato CorefUD poderia trazer benefícios práticos e, ainda, viabilizar o uso de modelos como os da família Sabiá, que não conseguiram responder no formato estabelecido.

Além disso, o modelo pode se beneficiar de aprimoramentos nos prompts utilizados ou até da adoção de outras abordagens de *prompting* - visando, especialmente, aprimorar recall dos modelos. Esse ponto é particularmente relevante para os prompts *few-shot*, que

apresentaram resultados inferiores à sua versão *zero-shot*, sugerindo que os exemplos utilizados não foram suficientemente eficazes e podem ser melhorados.

Outro ponto a ser considerado é a análise dos resultados dos modelos testados sob uma perspectiva mais linguística. Com a elaboração de novas ferramentas – ou a utilização de outros métodos – uma análise mais detalhada buscando correlações entre o desempenho dos modelos e variáveis linguísticas (como, por exemplo, a identificação da frequência de núcleos de diferentes classes gramaticais nos erros cometidos pelos modelos) pode possibilitar uma compreensão mais aprofundada de suas fraquezas e a elaboração de abordagens visando mitiga-las.

Ademais, dado que a revisão parcial do Summ-it++ se mostrou benéfica, seria interessante estender esse esforço ao corpus completo. Além disso, uma revisão similar no Coref-PT, atualmente o maior corpus com informações de correferência em português, pode ser uma boa contribuição aos recursos atualmente disponíveis a área de resolução e correferência no português.

Por último, os modelos podem ser testados utilizando textos com menções previamente identificadas, o que possibilitaria uma avaliação mais precisa de seu desempenho na tarefa central de resolução de correferência. Outra possibilidade seria explorar sua integração aos atuais modelos estado da arte em resolução de correferência.

REFERÊNCIAS

SALES ALMEIDA, Thales et al. **Sabiá-2: A New Generation of Portuguese Large Language Models**. arXiv e-prints, p. arXiv: 2403.09887, 2024.

ANIKINA, Tatiana; SKACHKOVA, Natalia; MOKHOVA, Anna. **Multilingual Coreference Resolution:** Adapt and Generate. EMNLP 2023, p. 19, 2023.

ANTONITSCH, André *et al.* **Summ-it++:** an enriched version of the Summ-it corpus. In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION, 10., 2016, Portorož. Proceedings [...]. Paris: European Language Resources Association (ELRA), 2016. p. 2047-2051.

BBN TECHNOLOGIES. **OntoNotes English Co-reference Guidelines**. Disponível em: <https://data.mendeley.com/datasets/zmycy7t9h9/2/files/722501a7-2c8e-435d-9920-a0ebe2081f2a>. Acesso em: 14 out. 2024.

COLLOVINI, S. *et al.* **Summ-it:** Um corpus anotado com informações discursivas visando a sumarização automática. In: CONGRESSO DA SBC, 27., 2007, Rio de Janeiro. Anais [...]. Rio de Janeiro: Sociedade Brasileira de Computação, 2007. p. 1605-1614.

CoNLL-U format. Disponível em: <https://universaldependencies.org/format.html>. Acesso em: 14 out. 2024.

CorrefVisual – PLN – PUCRS. Disponível em: <https://www.inf.pucrs.br/linatural/wordpress/recursos-e-ferramentas/correfvisual/>. Acesso em: 14 out. 2024.

CRUZ, André *et al.* Exploring *span*ish Corpora for Portuguese Coreference Resolution. In: **International Conference On Social Networks Analysis, Management And Security** (SNAMS), 5., 2018, Valencia. Proceedings [...]. Piscataway: IEEE, 2018. p. 290-295.

FONSECA, E. *et al.* CORP: Uma Abordagem Baseada em Regras e Conhecimento Semântico para a Resolução de Correferências. **Linguamática**, v. 9, p. 3-18, jul. 2017.

FONSECA, E. *et al.* Nominal Coreference Resolution Using Semantic Knowledge. In: COMPUTATIONAL PROCESSING OF THE PORTUGUESE LANGUAGE, PROPOR, 2018. Lecture Notes in Computer Science, Springer, v. 11122, p. 37-45, 2018.

FONSECA, Evandro *et al.* Improving coreference resolution with semantic knowledge. In: **International Conference Propor**, 12., 2016, Tomar. Proceedings [...]. [S.l.]: Springer International Publishing, 2016. p. 213-224.

FREITAS, Cláudia *et al.* Second HAREM: Advancing the State of the Art of Named Entity Recognition in Portuguese. In: **European Language Resources Association** (ELRA), 2010, Valletta, Malta. Proceedings [...]. Paris: European Language Resources Association, 2010.

GAN, Yujian *et al.* **Assessing the Capabilities of Large Language Models in Coreference:** An Evaluation. ACL Anthology, p. 1645–1665, maio 2024.

GARCIA, Marcos *et al*. Multilingual corpora with coreferential annotation of person entities. In: **International Conference On Language Resources And Evaluation**, 9., 2014, Reykjavik. Proceedings [...]. Reykjavik: European Language Resources Association (ELRA), 2014.

GITHUB. **CorefUD** *scorer*. Disponível em: <https://github.com/ufal/corefud-*scorer*>. Acesso em: 14 out. 2024.

HENDRYCKS, Dan *et al.* Measuring massive multitask language understanding. **arXiv preprint arXiv:2009.03300**, 2020.

HICKE, Rebecca MM; MIMNO, David. [Lions: 1] and [Tigers: 2] and [Bears: 3], Oh My! Literary Coreference Annotation with LLMs. arXiv preprint arXiv:2401.17922, 2024.

JURAFSKY, Daniel; MARTIN, James H. **Speech and Language Processing**. [S.l.: s.n.], 2024. Draft of August 20, 2024.

KNOTH, Nils *et al.* AI literacy and its implications for *prompt* engineering strategies. **Computers and Education: Artificial Intelligence**, v. 6, p. 100225, 2024.

LE, Nghia T.; BAI, Fan; RITTER, Alan. **Few-shot anaphora resolution in scientific protocols via mixtures of in-context experts**. arXiv preprint arXiv:2210.03690, 2022.

LE, Nghia T.; RITTER, Alan. **Are Large Language Models Robust Coreference Resolvers?**. arXiv preprint arXiv:2305.14489, 2023.

LIMA, Thiago *et al.* Analysing Semantic Resources for Coreference Resolution. In: VILLAVICENCIO, A. *et al.* Computational Processing of the Portuguese Language. PROPOR 2018. Lecture Notes in Computer Science, Springer, **Cham**, v. 11122, 2018.

Linguateca. Disponível em: <https://www.linguateca.pt/>. Acesso em: 14 out. 2024.

LIU, Chiu-Liang; HO, Chien-Ta; WU, Tzu-Chi. Custom GPTs Enhancing Performance and Evidence Compared with GPT-3.5, GPT-4, and GPT-4o? A Study on the Emergency Medicine Specialist Examination. In: **Healthcare**. MDPI, 2024. p. 172

LIU, R. *et al.* A brief survey on recent advances in coreference resolution. **Artificial Intelligence Review**, v. 56, n. 12, p. 14439–14481, maio 2023.

MARNEFFE, Marie-Catherine *et al.* Universal Dependencies. Computational Linguistics, v. 47, n. 2, p. 255–308, 2021.

MITKOV, Ruslan. **Anaphora Resolution. 1**. ed. Grã-Bretanha: Pearson Education, 2002.

MITKOV, Ruslan. **Anaphora resolution: the state of the art.** Wolverhampton, UK: School of Languages and European Studies, University of Wolverhampton, 1999.

MOREIRA, M. **Dunga detona Tite na Seleção: "Não coloquei minha mãe para dar entrevista"**. Disponível em: <https://www.cnnbrasil.com.br/esportes/futebol/dunga-detona-tite-na-selecao-nao-coloquei-minha-mae-para-dar-entrevista/>. Acesso em: 14 out. 2024.

NEDOLUZHKO, Anna. CorefUD 1.0: Coreference Meets Universal Dependencies. In: **Language Resources And Evaluation Conference**, 13., 2022, Marseille. Proceedings [...]. Marseille: European Language Resources Association, 2022. p. 4859–4872.

OPENAI. **ChatGPT (GPT-4).** Disponível em: <https://chat.openai.com/>. Acesso em: 14 out. 2024a.

OPENAI. **Hello GPT-4o.** Disponível em: <https://openai.com/index/hello-gpt-4o/>. Acesso em: 14 out. 2024b.

OPENAI. **Introducing GPTs.** Disponível em: <https://openai.com/index/introducing-gpts/>. Acesso em: 14 out. 2024c.

PAPE, David; EISENHOFER, Thorsten; SCHÖNHERR, Lea. *prompt* Obfuscation for Large Language Models. **arXiv preprint arXiv:2409.11026**, 2024.

POESIO, Massimo *et al.* **Anaphora Resolution:** Algorithms, Resources, and Applications. Berlim: Springer Nature, 2016.

PORADA, Ian; CHEUNG, Jackie Chi Kit. **Solving the Challenge Set without Solving the Task:** On Winograd Schemas as a Test of Pronominal Coreference Resolution. arXiv preprint arXiv:2410.09448, 2024.

PRADHAN, Sameer *et al.* CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In: **Joint Conference On Emnlp And CoNLL-Shared Task**, 2012. Proceedings [...]. Piscataway: IEEE, 2012. p. 1-40.

prompt ENGINEERING. ***prompt* engineering guide**. Disponível em: <http://www.*prompting*guide.ai>. Acesso em: 14 out. 2024.

PUSTEJOVSKY, James *et al.* **Natural Language Annotation for Machine Learning.** [S.l.]: O'Reilly, 2013.

RECASENS, Marta *et al*. Semeval-2010 task 1: Coreference resolution in multiple languages. In: **Proceedings of the 5th international workshop on semantic evaluation.** 2010. p. 1-8.

ROCHA, Gil *et al.* Towards a mention-pair model for coreference resolution in portuguese. In: **Epia Conference On Artificial Intelligence**, 18., 2017, Porto. Proceedings [...]. Porto: Springer International Publishing, 2017. p. 855-867.

ROZIERE, Baptiste *et al*. **Code llama:** Open foundation models for code. arXiv preprint arXiv:2308.12950, 2023.

RUSLAN, Mitkov. **Anaphora Resolution**. [S.l.]: Routledge, 2014.

SANTOS, Diana *et al.* HAREM: An Advanced NER Evaluation Contest for Portuguese. In: **European Language Resources Association**, 2006, Genoa. Proceedings [...]. Genoa: European Language Resources Association, 2006.

SAUSSURE, Ferdinand de. **Curso de linguística geral.** 27. ed. São Paulo: Cultrix, 2006.

SOON, M. *et al.* A machine learning approach to coreference resolution of noun phrases. **Computational Linguistics**, v. 27, n. 4, p. 521–544, 2001.

STRAKA, Milan *et al.* UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In: **Language Resources And Evaluation Conference** (LREC), 10., 2016, Portorož. Proceedings [...]. Paris: European Language Resources Association, 2016.

SUKTHANKER, Rhea *et al*. **Anaphora and coreference resolution:** A review. Information Fusion, v. 59, p. 139-162, 2020.

TAO, Guanhong *et al.* Opening a Pandora's box: things you should know in the era of custom GPTs. **arXiv preprint arXiv:2401.00905**, 2023.

VIEIRA, Renata *et al*. **Corref-PT:** A semi-automatic annotated portuguese coreference corpus. Computación y Sistemas, v. 22, n. 4, p. 1259-1267, 2018.

WILLIAMS, Alexander. In: **Stalmaszczyk P**, ed. The Cambridge Handbook of the Philosophy of Language. Cambridge Handbooks in Language and Linguistics. Cambridge University Press; p. 366-386, 2021.

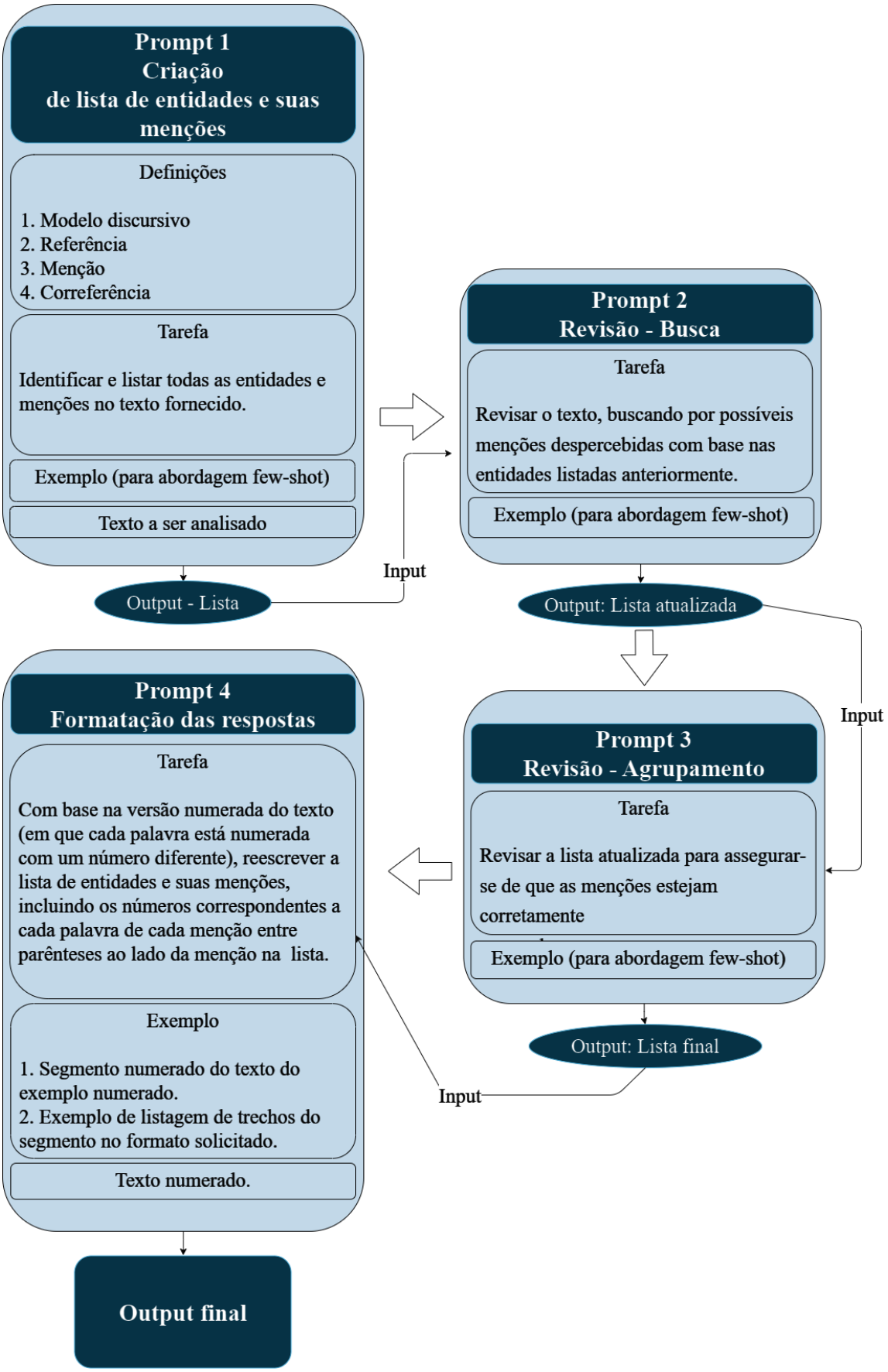
XU, Huiyu *et al.* RedAgent: Red Teaming Large Language Models with Context-aware Autonomous Language Agent. **arXiv preprint arXiv:2407.16667**, 2024.

YANG, Xiaohan *et al*. **What gpt knows about who is who. arXiv preprint arXiv:2205.07407**, 2022.

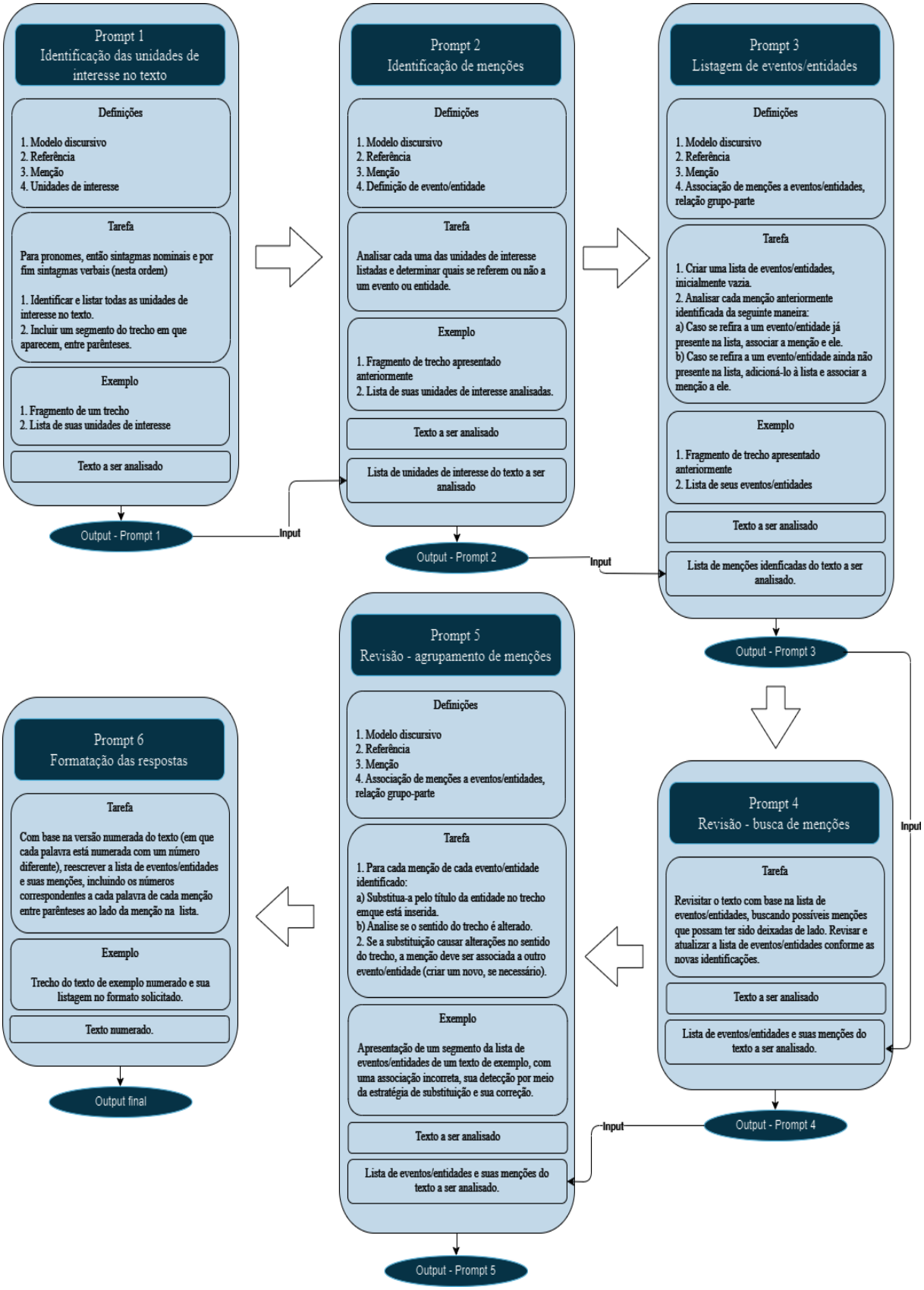
YU, Juntao *et al*. **The universal anaphora** *scorer***.** In: Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022. European Language Resources Association (ELRA), 2022. p. 4873-4883.

YU, Juntao *et al*. The universal anaphora *scorer* 2.0. In: **Proceedings of the 15th International Conference on Computational Semantics**. 2023. p. 183-194.

APÊNDICE A – ESTRUTURA DE *PROMPTS* ABORDAGEM A



APÊNDICE B – ESTRUTURA DE *PROMPTS* ABORDAGEM B



APÊNDICE C – PONTUAÇÃO DOS MODELOS TESTADOS

Resultados na resolução do segmento do corpus Summ-it++ original

Modelo	Abordagem	Média Prec	Média Recall	CoNLL	MUC			B^3			CEAFe			CEAFm			BLANC			LEA			MOR		
					P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
GPT4o	B	39,04	30,64	36.50	37,50	31,13	34,02	38,35	29,23	33,17	48,28	37,66	42,31	47,24	38,46	42,40	25,82	18,07	21,08	32,55	24,04	27,66	43,54	35,87	39,33
	Af	22,77	11,33	14,09	18,60	7,55	10,74	21,98	8,09	11,83	28,11	15,18	19,71	32,86	14,74	20,35	12,17	3,14	4,95	12,14	5,17	7,25	33,52	25,43	28,92
	A_0	34,52	8,81	15,34	30,77	7,55	12,12	37,21	7,54	12,53	42,08	14,31	21,35	48,84	13,46	21,11	24,89	3,22	5,63	27,91	4,74	8,11	29,94	10,87	15,95
GPT4c	B	37,11	34,39	37,41	38,78	34,91	36,74	36,44	32,22	34,20	41,28	41,28	41,28	46,62	44,23	45,39	25,30	22,10	23,59	30,11	26,24	28,04	41,22	39,78	40,49
	Af	30,20	30,23	32,03	31,00	29,95	30,10	31,33	28,41	29,80	34,25	38,36	36,19	38,46	38,46	38,46	21,72	18,15	19,72	24,00	22,19	23,06	30,63	36,09	33,13
	A_0	42,21	37,48	42,11	42,86	39,62	41,18	40,78	36,84	38,71	49,63	43,67	46,46	51,41	46,79	48,99	31,23	27,53	29,27	35,33	32,26	33,73	44,20	35,65	39,47

Resultados na resolução do segmento do corpus Summ-it++ revisado

Modelo	Abordagem	Média Prec	Média Recall	CoNLL	MUC			B^3			CEAFe			CEAFm			BLANC			LEA			MOR		
					P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
GPT4o	B	42,96	25,02	33,60	38,64	24,29	29,82	42,28	22,06	28,99	49,50	36,43	41,97	50,39	33,16	40,00	30,65	12,53	17,57	36,51	17,67	23,82	52,77	28,99	37,42
	Af	25,96	9,40	12,42	18,60	5,71	8,74	25,28	5,86	9,51	28,17	14,35	19,01	35,71	12,95	19,01	16,26	2,33	4,04	15,00	3,03	5,05	42,69	21,59	28,68
	A_0	42,74	8,17	14,17	42,31	7,86	13,25	43,60	5,94	10,45	38,71	12,42	18,80	53,49	11,92	19,49	32,66	2,44	4,50	38,14	4,45	7,96	50,30	12,17	19,60
GPT4c	B	42,18	28,54	34,38	40,82	28,57	33,61	41,57	24,42	30,77	39,91	37,65	38,75	47,97	36,79	41,64	35,13	18,14	23,92	35,16	19,01	24,68	54,73	35,22	42,86
	Af	41,53	29,66	34,59	40,00	27,86	32,84	41,51	23,44	29,96	39,86	42,11	40,95	48,08	38,86	42,98	31,58	15,69	20,95	35,09	16,77	22,69	54,61	42,90	48,05
	A_0	55,35	36,04	45,72	56,12	39,29	46,22	53,94	33,75	41,52	54,47	45,22	49,41	59,15	43,52	50,15	48,23	24,91	32,83	49,79	30,26	37,64	65,77	35,36	45,99