UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO

CENTRO DE CIÊNCIAS EXATAS, NATURAIS E DA SAÚDE

DEPARTAMENTO DE COMPUTAÇÃO

CIÊNCIA DA COMPUTAÇÃO

Gabriel Magalhães Dias

Extração de Relacionamentos entre Entidades Nomeadas

ALEGRE - ES

JULHO DE 2022

Extração de Relacionamentos entre Entidades Nomeadas

Trabalho de conclusão de curso apresentado ao Departamento de Computação do Centro de Ciências Exatas, Naturais e da Saúde da Universidade Federal do Espírito Santo, como requisito parcial para obtenção do grau de Bacharel em Ciência da Computação.

por

Gabriel Magalhães Dias

Orientadora
JULIANA PINHEIRO CAMPOS PIROVANI

Universidade Federal do Espírito Santo

ALEGRE - ES

JULHO DE 2022

GABRIEL MAGALHÃES DIAS

Extração de Relacionamentos entre Entidades Nomeadas

Trabalho de Conclusão de Curso apresentado ao Departamento de Computação do Centro de Ciências Exatas, Naturais e da Saúde da Universidade Federal do Espírito Santo como requisito parcial para obtenção de grau de Bacharel em Ciência da computação.

Aprovado em 11 de agosto de 2022.

COMISSÃO EXAMINADORA

Prof. Drª, Juliana Pinheiro Campos Pirovani
Universidade Federal do Espírito Santo
Orientadora

Prof. Dr. Edmar Hell Kampke
Universidade Federal do Espírito Santo

Prof.ª. MSc. Valéria Alves da Silva
Universidade Federal do Espírito Santo

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

PLN	Processamento de Linguagem Natural
REN	Reconhecimento de Entidades Nomeadas
EI	Extração de Informação
EN	Entidades Nomeadas
ER	Extração de Relações
GL	Gramáticas Locais
CD	Coleção Dourada

RESUMO

O Reconhecimento de Entidades Nomeadas tem como objetivo identificar e classificar automaticamente entidades como pessoas, locais e organizações em textos não estruturados. A Extração de Relações, por sua vez, é a tarefa voltada para extrair relações entre conceitos como, por exemplo, duas entidades nomeadas. A Extração de Relações é uma tarefa muito importante para a Extração de Informação. Entretanto, se trata de uma tarefa complexa. Para realizá-la, podem ser utilizadas algumas abordagens. Neste trabalho, foi utilizada a abordagem linguística adicionando Transdutores, que são os grafos definidos dentro de uma Cascata de Transdutores, essa que é uma coleção ordenada desses grafos. A Cascata que foi trabalhada tem o objetivo de extrair essas relações e ela foi construída em um trabalho anterior. O objetivo deste trabalho foi a identificação de relações entre Entidades Nomeadas. Para tal, foi usada a ferramenta Unitex, a mesma utilizada para elaborar a Cascata que servirá de base. Os scripts de avaliação e os *corpora* do HAREM, um evento da Linguateca, foram usados para avaliação de desempenho da Cascata de Transdutores construída para extrair relações. Como resultado, obtivemos uma precisão de 66,7% nas relações extraídas.

Palavras-chave: Reconhecimento de Entidades Nomeadas; Extração de Relações; Unitex; Cascata de Transdutores.

1 INTRODUÇÃO

Na era moderna, com a transição de livros e outros documentos impressos, que estão escritos utilizando linguagem natural, para o formato digital, surgiu a possibilidade e a necessidade do computador interpretar esses conhecimentos. Como esses textos em linguagem natural não são compreendidos pela máquina, o Processamento de Linguagem Natural (PLN) surge como uma importante subárea da Computação para alcançar este objetivo: Fazer com que a máquina possa trabalhar com as informações contidas nesses textos. Hoje, observamos o PLN sendo muito utilizado em *chatbots* ou alguns *softwares* cujo objetivo é a interação com um humano através de textos ou áudio, mas de modo não "pré-programado", ou seja, sem as entradas com cenários previsíveis e bem definidos, mas com uma conversa interativa e sem um controle tão rígido pelo *software*.

O Reconhecimento de Entidades Nomeadas (REN) é uma tarefa das áreas de PLN e Extração de Informação (EI). A mesma tem como objetivo identificar e classificar entidades automaticamente em textos de escrita livre, variando entre nome de pessoas, lugares e organizações, até obras e datas, dependendo da relevância em certos domínios. Tais informações podem ser úteis para saber a quem o texto se refere, onde e quando algo aconteceu, melhorando a compreensão do mesmo. Porém, o REN não é uma tarefa simples. Várias categorias de Entidades Nomeadas (EN) são escritas de forma semelhante e aparecem em contextos semelhantes. Por exemplo, nomes de pessoas e lugares começam com uma letra maiúscula, assim como expressões temporais e valores contém números. Além disso, a mesma EN pode ser classificada em categorias diferentes dependendo do contexto em que aparece: a EN Washington pode se referir a uma pessoa em um contexto e a um local em outro.

Segundo Pirovani e Oliveira (2015), o REN depende do idioma, do *corpus*[1] e do domínio, ou seja, do contexto do qual ele foi escrito. Considerando a dependência do domínio, a mesma categoria de EN pode ser escrita de formas diferentes em diferen-

[1] Corpus é um conjunto de textos

tes gêneros textuais. Por exemplo, em textos de e-mails é comum aparecer nomes de pessoas após palavras como *Olá* e *Boa tarde*, enquanto em textos de portarias e memorandos é comum aparecer nomes de pessoas após palavras como *servidor* e *professor*.

A Extração de Relações (ER) também é uma tarefa da EI que é responsável pela descoberta de relacionamentos semânticos entre conceitos em textos não estruturados (SOUZA; CLARO, 2014). Neste trabalho, ela consiste em detectar a relação entre um par de ENs, considerando um determinado texto.

Sistemas de ER podem ser desenvolvidos utilizando as seguintes abordagens: linguística, aprendizado de máquina ou híbrida. Na abordagem linguística, regras onde relações podem aparecer são identificadas e construídas manualmente, ou seja, as regras permitem a detecção de relações. No aprendizado de máquina, sistemas aprendem a identificar e classificar relações a partir de um *corpus* de treinamento. A abordagem híbrida combina as duas abordagens citadas. Uma forma de representar as regras da abordagem linguística são as Gramáticas Locais (GL), formalismo introduzido por Gross (1997), ou Cascata de Transdutores, utilizado por Friburger e Maurel (2004).

No trabalho de Magalhaes e Pirovani (2020), foi desenvolvida uma Cascata de Transdutores, a partir da GL construída em Pirovani (2019), para o REN em português. Com essa versão estabelecida sobre as categorias obtidas, este trabalho se voltou para obter as relações entre essas entidades encontradas, assim conseguindo obter mais informações sobre um determinado texto.

1.1 Problema e sua Importância

Segundo Abreu (2014), "por vezes o Reconhecimento de Entidades Nomeadas não é suficiente para tarefas de EI, requerendo também a identificação das relações estabelecidas entre essas entidades".

Tal ponto também foi reforçado devido a complexidade em obter certas entidades em alguns contextos. Determinar a relação entre as que conseguem ser marcadas mais facilmente com as de difícil contexto pode futuramente potencializar os resultados. Além da capacidade de obter uma maior abrangência de entidades marcadas, a ER proporciona um entendimento melhor sobre o texto, entendendo onde duas ENs possuem o mesmo significado, onde uma se "inclui" em outra e outros tipos de relações

que serão abordadas na Seção 2.

Uma das maiores dificuldades é que a abordagem linguística, assim como as outras, depende diretamente da linguagem em questão e o Português, atualmente, não se encontra tão favorecido como o Inglês ou outras línguas de países pioneiros na área de PLN. Esses países já possuem mais ferramentas desenvolvidas, porém, por essa dependência da gramática e outras peculiaridades da fala e da escrita do idioma, a conversão dessas ferramentas é um objetivo complexo.

No trabalho desenvolvido por Pirovani (2019), a Gramática Local apresentava bons resultados mas, com a adição de regras, como vistas em Magalhaes e Pirovani (2020), pôde ser aprimorada e alcançou melhores resultados em distintos contextos e nos *corpus* usados como base de estudo. Para tal objetivo, como foi citado previamente, a extração de relações se mostra como uma boa opção para alcançar mais ENs. Porém, essa tarefa se mostra complexa de início já que existem diversos tipos de relacionamentos e, além disso, elas podem não ser tão diretas como desejável. Assim, o foco será bem definido em cinco classificações extraídas do SegundoHaremRelRelem, um dos *corpus* disponibilizados pela Linguateca e que será detalhado na seção 2. Essas categorias são: identidade, inclusão, ocorre e sede em, relações familiares e da categoria outros. Essas relações também serão abordadas com mais detalhes na seção 2.1.

A importância deste trabalho é ressaltada devida uma necessidade do PLN no cenário atual. Isso por conta da evolução de softwares de *chatbots*, assistentes virtuais e outros como filtros de pesquisa, como o *Google* ou outros buscadores, tradução de idiomas, análises de dados e de textos a fim de recuperar as informações contidas no documento digital e etc. Além dos *softwares* de interação com humanos, a recuperação de informações de Redes Sociais como *Twitter* se mostra extremamente relevante no contexto atual, já que as Redes Sociais se mostram como a mais efetiva forma de comunicação e transmissão de informação. Assim, a evolução dessas tarefas é relevante. A ER pode contribuir na extração de certas informações contidas nos textos que são utilizadas pelas empresas citadas e demais arquivos, documentos históricos e os gerados atualmente.

Com tais observações, se mostra relevante a abordagem da ER com Cascata de Transdutores. Ela pode ser usada tanto de forma individual, quanto contribuindo com outras abordagens como a de aprendizado de máquina, assim como apresentado no trabalho de REN de Pirovani (2019) e Magalhaes e Pirovani (2020).

1.2 Objetivos

1.2.1 Objetivo Geral

Construir uma Cascata de Transdutores para extrair relações entre entidades nomeadas considerando as categorias pessoa, local e organização, a partir da Cascata de Transdutores de Magalhaes e Pirovani (2020) que reconhece EN.

1.2.2 Objetivos específicos

1. Estudar e compreender como os *corpora* de referência para o português anotam as relações entre entidades.
2. Avaliar os resultados atuais da Cascata de Transdutores.
3. Analisar os padrões existentes para ER entre as categorias escolhidas nos *corpora* de referência para o português, de modo a estar apto a descrever esses padrões.
4. Inserir regras que capturem os padrões identificados na Cascata de Transdutores de Magalhaes e Pirovani (2020).
5. Aplicar a Cascata de Transdutores nos *corpora* de teste.
6. Avaliar os resultados obtidos.

2 REVISÃO DE LITERATURA

2.1 Eventos de avaliação de Extração de Informação

Para obter alguns *corpus* para o estudo em questão, foram vistos alguns eventos que tem o objetivo de anotar ou fornecer *corpora* (*corpus* ou conjuntos de textos). Alguns desses serão abordados com mais detalhes.

O HAREM (SANTOS; CARDOSO, 2007) foi um evento organizado pela Linguateca (LINGUATECA, 2019). Ele foi um grande incentivador para o REN na língua portuguesa. Nele as ENs são classificadas em 10 categorias, sendo elas: Abstração, Acontecimento, Coisa, Local, Obra, Organização, Pessoa, Tempo, Valor e Outro. Além disso, ele disponibilizou *corpora* anotados, ou seja, com as entidades nomeadas relacionadas a essas categorias já identificadas e classificadas. Esses *corpus* são conhecidos como Coleção Dourada (CD) e são divididos em Primeiro HAREM, Mini HAREM e Segundo HAREM, de acordo com os textos disponibilizados em cada evento.

O SegundoHarem[1], que é a última coleção lançada, é formado por 129 textos tanto escritos em Português do Brasil quanto o nativo (de Portugal). Sua importância se dá por ser o primeiro onde a extração de relacionamentos entre entidades nomeadas foi explorada, fornecendo *corpus* anotado e definindo algumas classificações para as relações. Elas se dividem em cinco:

1. Identidade: Que se trata de duas ou mais ENs que compartilham o mesmo significado, tais como siglas, por exemplo. Para ilustrar, temos a sentença: "A Universidade Federal do Espírito Santos (UFES)", onde "Universidade Federal do Espírito Santo" tem o mesmo significado que "UFES".

2. Inclusão: Onde um elemento A está incluso em um elemento B, podendo ser fisicamente como um estado em um país, ou por definição, como um ano em uma

[1] Corpus do HAREM disponível em: www.linguateca.pt/HAREM/

década. Na sentença: "Centro de Convenções de Curitiba, endereço presente há muitos anos na cidade, escondido na Rua Barão do Rio Branco.", temos essa relação, onde o "Centro de Convenções de Curitiba" se localiza dentro da "Rua Barão do Rio Branco".

3. "Ocorre em" ou "Sede de": Sugerem as relações entre entidades que citam um evento ou empresa que tem sede em determinado local. No texto: "Em 9 de Setembro de 1895, foi organizado em New York o Congresso Americano de *Bowling*", a relação se dá entre a entidade "Congresso Americano de Bowling" que foi realizado em "New York".

4. Relação Familiar: São relações que envolvem duas pessoas que são familiares, como pais e filhos, avós e netos e outras. Exemplo: "Carlos é pai de Pedro".

5. Outras: Uma classificação mais genérica foi definida para relações que não se enquadram nas classificações anteriores. Um exemplo da mesma é a relação entre um funcionário e uma empresa. Como o exemplo: "Ricardo, funcionário da AutoGlass".

Para representar as relações no *corpus* em questão, é utilizado um padrão baseado em *tags*. A Figura 1 apresenta um exemplo dessa anotação. Onde a EN *Reinaldo Machado* está marcada no padrão XML do HAREM. Nota-se que o atributo *ID* contém um identificador **único** para a entidade marcada, *CATEG* contém a categoria correspondente entre as 10 classificadas pelo HAREM.

```
<EM ID="ric-46546-14000" CATEG="PESSOA">Reinaldo Machado</EM>
```

Figura 1: Exemplo de anotação de EN no padrão do HAREM

Para relacionar uma EN com outra, o *COREL*, mostrado na Figura 2, é indicado na entidade que foi marcada devido a relação. Dessa forma, podemos saber qual entidade está relacionada àquela em questão. Assim, temos o *ID* da entidade "origem da relação" no atributo *COREL*. Já o atributo *TIPOREL* é aquele que indica qual dos tipos de relacionamento é o cenário em questão, variando entre *relacao_familiar* para relações de parentesco, *ident* para identidade, *inclui* e *incluido* para inclusão, *ocorre_em* e *sede_de* para as relações de "ocorre em" e, por fim, *outra* para qualquer relação que não corresponde as classificações anteriormente citadas. A Figura 2 exemplifica uma anotação de relação.

```
    <EM ID="ric-46546-14000" CATEG="PESSOA">Reinaldo Machado</EM>, cujo pai,
<EM ID="ric-46546-14001" CATEG="PESSOA" TIPOREL="relacao_familiar" COREL="ric-46546-14000">José Alves Machado</EM>
```

Figura 2: Exemplo de anotação de relações no padrão do HAREM

Além desses *corpus*, foram obtidos os *corpus* já *tokenizados* do IberLEF (COLLOVINI et al., 2019). Esse evento também é relevante na área de extração de informação de textos de linguagem ibéricas. A importância deles nesse trabalho se dá nos estudos de relação, com textos que também foram marcados, mas de forma mais específica.

Neste evento, as relações devem ser marcadas como uma tripla, assim como no seguinte exemplo: "A Marfinite fica em o Brasil.", deve ser marcado "(Marfinite, fica em, Brasil)". Essa tripla envolve as duas EN marcadas e a relação entre elas. Nesse *corpus* só são envolvidas as classificações de relações entre Pessoa, Local e Organização.

2.2 Unitex, Gramática Local e Cascata de Transdutores

Além dos estudos dos *corpus*, foi necessária a revisão nos temas relevantes para o contexto deste trabalho, sendo estes: Unitex (MUNIZ; NUNES; LAPORTE, 2005), Gramática Local (GROSS, 1999) e Cascata de Transdutores (FRIBURGER; MAUREL, 2004).

2.2.1 Unitex

O Unitex é um conjunto de softwares livres para a construção e aplicação de gramáticas e cascatas. Ele foi escolhido para adaptar a GL, pois foi utilizado anteriormente por (PIROVANI, 2019), e principalmente pela facilidade de gerar gramáticas locais e cascatas de transdutores. Nas próximas seções será abordado como ambas são geradas na ferramenta.

2.2.2 Gramática Local

Uma gramática local é um autômato com uma sequência de estados no qual podemos definir uma regra. No nosso caso, desejamos uma regra linguística no idioma

português. Logo, utilizamos a GL para definir manualmente uma regra que identifica determinada entidade nomeada ou relação.

> ”Gramáticas locais são gramáticas de estados finitos ou autômatos de estados finitos que representam conjuntos de expressões de uma língua natural”, Gross (1999).

No Unitex, uma Gramática Local pode ser representada como um ou mais grafos, como a apresentado na Figura 3. Na GL de (PIROVANI, 2019), que serviu de base para este trabalho, cada uma das 10 categorias do HAREM possuía, pelo menos, um grafo para identificar e anotar ENs da categoria.

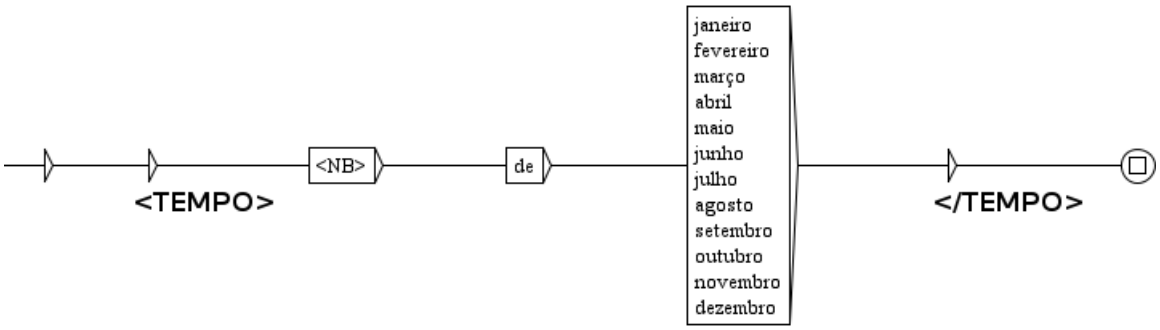

Figura 3: Regra do grafo que identifica ENs da classe TEMPO.

A regra do grafo apresentado na Figura 3 identifica ENs da classe Tempo. Esse grafo reconhece números (identificados pelo código *<NB>* nos dicionários do Unitex) seguidos pela preposição "*de*" e, após essa preposição, palavras que representam meses do ano como "*janeiro*" ou "*dezembro*". Símbolos que aparecem entre < e > são interpretados pelo Unitex como código de propriedade lexical nos dicionários ou como lema.

O Unitex permite incluir saídas no grafo, representadas em negrito sob setas. Ao aplicar grafos para extrair padrões em um texto, as saídas podem ser anexadas ao arquivo de concordância (opção *MERGE with input text*). Assim, as ENs são anotadas pelo Unitex com *tags* no início e no fim de acordo com a categoria em questão (*<TEMPO>* e *</TEMPO>* neste exemplo).

Exemplos de entidades que serão marcadas pelo grafo da Figura 3 são:
“*<TEMPO>* 15 de Novembro *</TEMPO>*”
“*<TEMPO>* 26 de abril *</TEMPO>*”.

2.2.3 Cascata de Transdutores

Uma Cascata de Transdutores é um conjunto de Gramáticas Locais com saídas que podem ser aplicadas em uma determinada ordem de forma que as últimas utilizem os resultados das primeiras em suas regras.

A Cascata aplica vários grafos com saídas (transdutores), um após o outro, em um texto: cada grafo modifica o texto e as alterações podem ser úteis para processamentos posteriores com os próximos grafos (PAUMIER, 2021). Portanto, é necessário definir uma ordem específica para aplicar os grafos em sequência aos textos de entrada.

Com essa definição apresentada, a decisão de usar uma Cascata de Transdutores se dá pela ambiguidade que a Gramática Local tem ao chamar outros grafos. Com uma ordem pré-definida, podemos ter um maior controle sobre o uso dos resultados de um subgrafo ou de um grafo focado em outra categoria em um próximo.

Existe uma forma especial para anotar padrões usando Cascata de Transdutores no Unitex. A partir do momento que reconhecemos um padrão (como xxx) e o anotamos com uma *tag* lexical ({xxx,. TEMPO}), uma máscara lexical (<*TEMPO*>) poderá ser usada para reconhecer aquele padrão em outros grafos posteriormente.

Um caso para ilustrar foi o ocorrido no trabalho de Magalhaes e Pirovani (2020), onde foi observado que algumas datas (classificadas previamente como TEMPO) apareciam como nomes de locais (Ex: Rua 15 de Novembro). Usando a *tag* lexical na GL correspondente para anotar ENs da categoria Tempo, essas ENs poderiam ser usadas posteriormente no reconhecimento de ENs da categoria Local. Nesse caso, a GL que reconhece Tempo deve ser aplicada antes da que reconhece Local.

A Figura 4 apresenta o grafo que também identifica as ENs da categoria de TEMPO, adaptado com a *tag* lexical para a Cascata de Transdutores.

A Figura 5 apresenta uma regra da Cascata de Transdutores que identifica ENs da classe Local utilizando ENs previamente reconhecidas como Tempo.

Exemplos de entidades que serão anotadas pela regra da Figura 5 são:
“<*LOCAL*> Praça 7 de setembro </*LOCAL*>”
“<*LOCAL*> Rua 15 de Novembro </*LOCAL*>”.

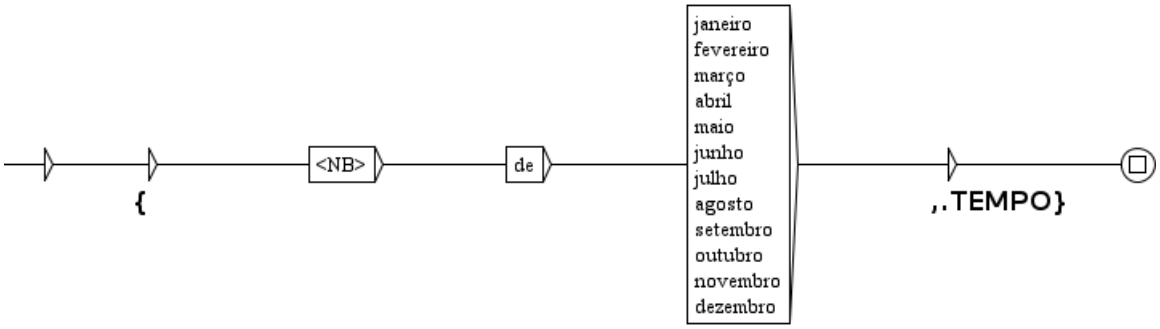

Figura 4: Grafo correspondente ao da Figura 3, porém adaptado para a Cascata de Transdutores.

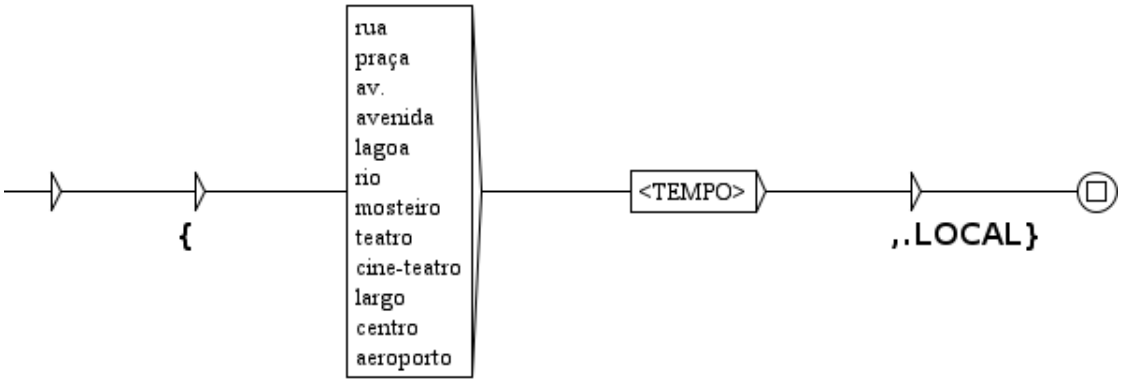

Figura 5: Regra da Cascata de Transdutores que identifica ENs da classe LOCAL.

2.3 Trabalhos Correlatos

Este trabalho tem como principal referência o trabalho de Magalhaes e Pirovani (2020) que teve como objetivo, melhorar os resultados de Pirovani (2019). Para a GL desenvolvida em Pirovani (2019), foram estudadas regras em textos de diferentes gêneros textuais como e-mails, portarias e entrevistas. Com essa detecção de novas regras, foram construídos grafos referente as mesmas com a finalidade de marcar as 10 diferentes classificações de entidades do HAREM, citadas anteriormente.

A partir dessa versão melhorada da GL base, uma nova Cascata foi construída, utilizando os grafos de forma adaptada, com a tarefa de conseguir melhores resultados. Tal análise foi feita nos *corpus* que também serão utilizados neste trabalho: PrimeiroHarem, MiniHarem e SegundoHarem. Os resultados finais da Cascata de Transdutores de Magalhaes e Pirovani (2020) estão na tabela 1.

> Ao adaptar a GL para Cascata foi notada a clara diferença entre o funcionamento de ambas, percebendo o aumento significativo da Cascata. Com as alterações e a utilização adequada da estrutura, a Cascata apresentou um melhor desempenho por causa da utilização de

> resultados prévios, adicionando contexto em algumas situações, aprimorando regras e impedindo ambiguidade em outras. (MAGALHAES; PIROVANI, 2020).

A Cascata resultante do trabalho de Magalhaes e Pirovani (2020) serviu como base para este, a fim de continuar o processo de aperfeiçoamento.

Corpus	Precisão	Abrangência	Medida-F
PrimeiroHarem	0.727	0.499	0.592
MiniHarem	0.726	0.485	0.582
SegundoHarem	0.717	0.446	0.550

Tabela 1: Resultados Finais da Cascata do trabalho Magalhaes e Pirovani (2020)

Além dos trabalhos que servirão de base, o trabalho de Collovini et al. (2019), faz um *overview* sobre reconhecimento de relações na língua portuguesa, baseada nas *tasks* 2 e 3 do IberLEF 2019. Elas são importantes nesse contexto já que a *task* 2 se trata da extração automática das relações entre PESSOA, LOCAL e ORGANIZAÇÃO. Essas são 3 das 10 categorias do HAREM e que são envolvidas diretamente nas cinco categorias de relações que trabalhamos. Além dessa definição, é explorada a relação entre 2 entidades de categorias diferentes, tais como o exemplo citado: "Roberto Lemos, diretor da Creative Commons", mostrando um relacionamento entre uma Pessoa e uma Organização.

A *task* 3 aborda a extração de relações "abertas", em outras palavras, de formas não estruturada. Isso aumenta muito o nível da complexidade do trabalho e, para obter melhores resultados, foi optado por seguir apenas a *task* 2.

O trabalho de Souza e Claro (2014) tenta mapear algumas ferramentas do inglês para a língua portuguesa. Além disso, este trabalho cita algumas definições importante como a de relações específicas, ou seja, aquelas onde há uma relação declarada, assim como as que são abordadas pelo HAREM e as relações abertas, que são as que não se limitam ao vocabulário presente no *corpus*, em outras palavras, uma relação que não necessariamente será declarada de forma explícita no texto. Isso permite que sejam marcadas muito mais entidades, porém a ambiguidade e a falta de contexto se mostra um empecilho. Portanto, será trabalhado apenas as relações específicas. Outro ponto importante notado nesse trabalho foram os resultados da tentativa do mapeamento, onde algumas *features* não apresentam um resultado satisfatório, com a *medida-F*, a métrica de resumo, abaixo dos 20%.

3 METODOLOGIA

O primeiro passo para elaborar as regras que formarão os novos Transdutores foi a análise dos *corpora* disponíveis em português, fornecidos pelo HAREM. Esse estudo envolveu observar os padrões do contexto das relações e a forma em que são anotadas. Esse estudo será voltado exclusivamente para o sub-grupo de relações das classificações definidas. Em seguida, as regras para anotarem essas relações foram adicionadas na Cascata de Transdutores obtidas do trabalho de Magalhaes e Pirovani (2020), utilizando a ferramenta do Unitex e por fim, após obter os resultados finais, foram levantadas as estatísticas sobre acertos pelas métricas de precisão, abrangência e medida-F.

Um dos principais pontos a serem esclarecidos se dá na ferramenta que foi escolhida, o Unitex. Sua capacidade para criar grafos mais complexos de maneira simples se mostrou muito intuitiva e de fácil leitura. Outro ponto que a favorece é a sua capacidade de marcação, elaborar Cascatas de Transdutores e, principalmente, a utilização de suas funcionalidades via linha de comando. Tal meio de executar essa ferramenta permitiu a elaboração de *scripts* com os quais foram anotam as relações automaticamente e calcula os resultados de acordo com o que se deseja para um determinado *corpus*.

Para avaliar os resultados obtidos, o *corpus* do HAREM foi utilizado como referência, devido a alguns pontos como: O padrão bem explícito de anotação, o volume de textos, diversidade de cenários e classificações. Já o *corpus* do IberLEF foi utilizado apenas como uma base de estudos. Dentre os *corpus* disponibilizados no HAREM, o **SegundoHaremRelRelem** foi o escolhido como parâmetro de avaliação.

O SegundoHaremRelRelem é uma versão do SegundoHarem onde a anotação também considera as relações, trabalhando no mesmo conjunto de textos. Ao todo, ele apresenta 129 textos que variam desde documentos históricos até artigos de jornais e revistas. Também possui cerca de 7.846 entidades nomeadas anotadas e dentre

elas, 3.784 relações foram etiquetadas. Dentre essas 3.784 relações, muitas delas foram anotadas considerando o texto como um todo, incluindo relações entre sentenças diferentes, o que é uma tarefa simples para um humano, mas se torna uma tarefa complexa para a máquina. A ferramenta Unitex é limitada a anotação dentro de uma sentença, assim impedindo a análise de todo um parágrafo ou do texto por completo.

Considerando isso, foi feita uma classificação, entre relações diretas e indiretas. As diretas são aquelas que são definidas dentro de uma sentença e as indiretas, são as que dependem do contexto do texto como um todo. Neste trabalho, focamos a criação das regras somente para as diretas. Porém, os resultados dos *scripts* trabalham com uma visão geral do texto, considerando todas as relações e não somente as diretas, e isso prejudica os resultados gerais.

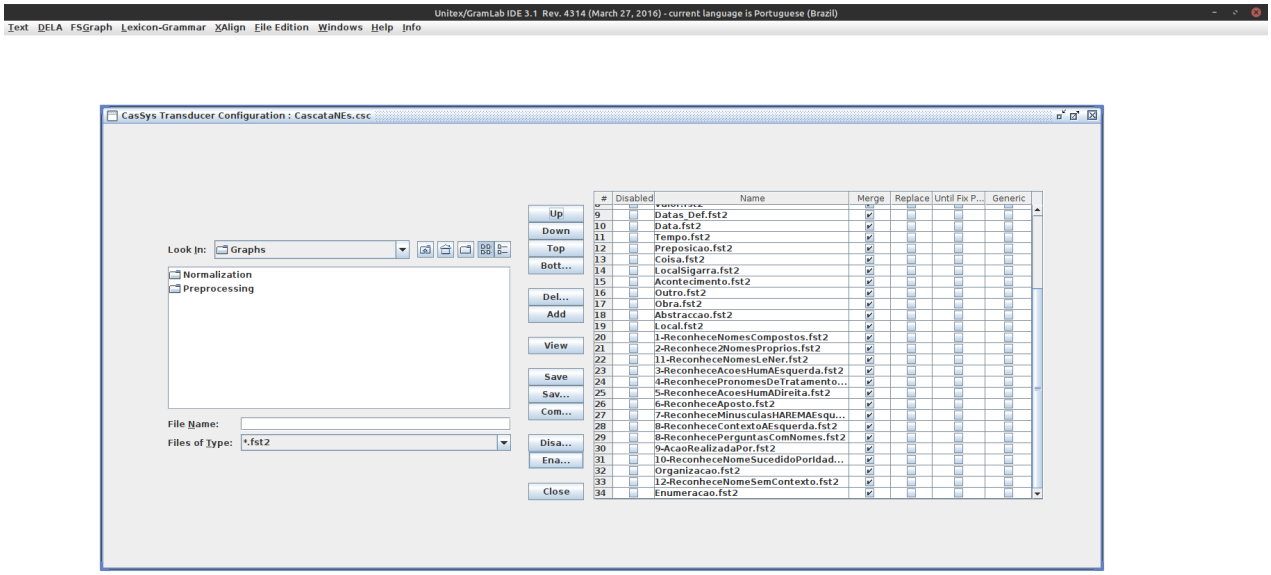

Figura 6: A cascata que servirá como base para este trabalho

Na Figura 6 vemos a Cascata de Transdutores do trabalho de Magalhaes e Pirovani (2020), que anota as 10 classificações do HAREM. Como a tarefa deste trabalho é a marcação de PESSOA, ORGANIZAÇÃO e LOCAL, as outras sete ainda são executadas pelo Unitex mas ignoradas na marcação, com o objetivo de avaliar somente o desempenho da extração de relações. Como dito acima, isso pode ser configurado pelos *scripts* elaborados. Para anotar e obter os resultados nos corpus da CD do HAREM, um conjunto de *scripts* que foi elaborado e publicado no próprio evento, foi utilizado. Os *scripts* aplicam a Cascata de Transdutores em um conjunto de textos e compara com um texto já etiquetado por um humano, como um "gabarito".

As métricas utilizadas seguem o padrão estabelecido em Pirovani (2019) e Magalhaes e Pirovani (2020). A precisão, abrangência e medida-F da Cascata, sendo a Precisão a porcentagem de acerto na marcação das entidades em comparação ao total marcado pela Cascata, a abrangência que calcula a porcentagem de acerto na marcação das entidades em comparação ao corpus original marcado por um humano e a medida-F que é a métrica de resumo, fazendo um calculo baseado nas duas citadas. O último resultado obtido pela Cascata em Magalhaes e Pirovani (2020) considerando as 10 entidades do HAREM é apresentado na Tabela 1

$$\textit{Precisão} = \frac{Total\ de\ ENs\ identificadas\ corretamente}{Total\ de\ ENs\ identificadas} \tag{6}$$

$$\textit{Abrangência} = \frac{Total\ de\ ENs\ identificadas\ corretamente}{Total\ de\ ENs\ realmente\ existentes\ no\ corpus} \tag{6}$$

$$\textit{Medida-F} = \frac{2\ *\ \textit{Precisão}\ *\ \text{Abrangência}}{\textit{Precisão}\ *\ \textit{Abrangência}} \tag{6}$$

Considerando a explicação sobre as relações diretas e indiretas e que os *scripts* entregam os resultados considerando todas as relações, a métrica que deve ser o destaque nos resultados é a **precisão**. Ela é a que representa a média de acertos considerando as anotadas pela gramática. Já as métricas de abrangência e *medida-F* apresentam resultados menores, devido a essa visão que considera as relações indiretas e essas não foram trabalhadas devido as limitações que também já foram citadas.

4 RESULTADOS

Para detalhar os resultados, as principais regras adicionadas serão apresentadas a seguir por tipos de relações e entidades envolvidas, ficando assim mais claro os impactos de cada uma. Inicialmente, trabalhamos com as relações de parentesco.

4.1 Relações de Parentesco

As relações de parentesco são aquelas que envolvem duas pessoas que possuem uma relação familiar, de sangue como pai e filho, ou não, como um padrasto e um enteado. Por exemplo, na seguinte sentença: "Rodrigo é neto de Sebastião e pai de Carlos", temos as relações entre Rodrigo e Sebastião e Rodrigo com Carlos, onde a primeira corresponde a relação de um avô e seu neto e a segunda de um pai e filho.

Para trabalhar inicialmente com as relações foi necessário separar as antigas regras, que trabalham com etiquetagem de nomes, das novas, que trabalham com etiquetagem de relações, assim tendo uma visão mais separada de cada tarefa. Para tal, foi dividido o grafo *PalavrasMinusculasHarem.grf* em um novo contendo apenas as palavras de relacionamentos familiares como irmão, pai, padastro, primo, tio, sobrinho e suas flexões para feminino, plural e demais formas de nomear relações familiares, como o caso de avô e vovô.

Com esse novo grafo poderíamos diferenciar as demais palavras que se referem a profissões, pronomes de tratamento e demais casos das que se referem a uma posição na família. Por manter uma responsabilidade parecida com o anterior, ele foi nomeado como *PalavrasMinusculasFamilia.grf* e foi definido como podemos ver na Figura 7.

Para fazer o estudo para a adição de novas regras, foi utilizado o *corpus* do SegundoHarem considerando a entidade de origem e alvo. Para relação familiar, é necessário que ambas entidades envolvidas sejam da categoria PESSOA. Com esse filtro, encontramos 822 ocorrências dentro do *corpus*. Para reduzir e ser mais objetivo na

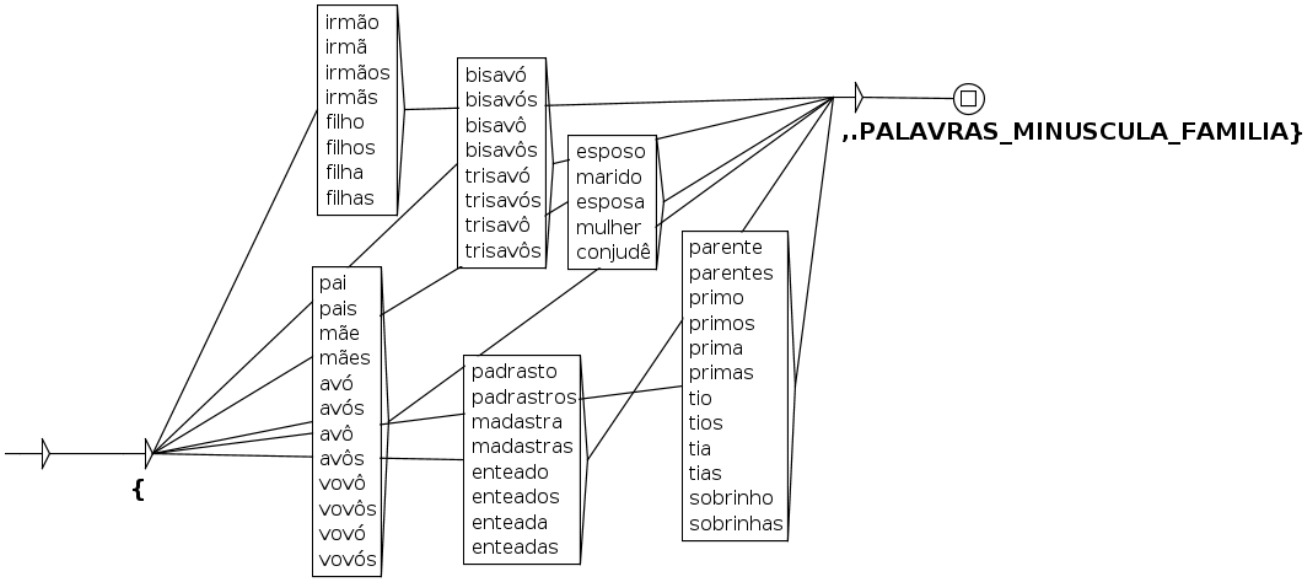

Figura 7: Grafo responsável por encontrar palavras associadas a certa relação familiar

categoria, foram separadas dentro dessas 822, apenas aquelas onde o tipo de relação era classificada como "relacao_familiar" e assim chegamos a 32 casos.

Corpus	Precisão	Abrangência	Medida-F
SegundoHarem	0.828	0.010	0.002

Tabela 2: Resultados da primeira versão da anotação de relações familiares

Como podemos observar na Tabela 2, o resultado de precisão ficou alto e as demais, ficaram baixas. Isso se dá porque neste trabalho, foram anotadas apenas as relações mais objetivas, definidas como relações diretas, que são minoria dentro do *corpus*, assim como foi explicado no início da sessão. Nas demais tabelas, serão apresentadas apenas o resultado da precisão.

Analisando o resultado em si da Tabela 2, foi notado que a versão da Cascata com as relações encontrou 9 relações familiares, sendo duas delas, consideradas erradas. A mesma se dá na sentença apresentada na Figura 8, onde o erro é compreensível devido ao contexto da mesma. Tal situação demanda um estudo mais aprofundado do idioma e de sua escrita para determinar se Zeca Afonso poderia ser, de fato, filho da Madrugada.

Também é importante ressaltar que para obter os resultados detalhados para as relações, foram necessários ajustes nos *scripts* utilizados como base, desenvolvidos no trabalho de Magalhaes e Pirovani (2020). Essas modificações foram realizadas no

```
no tributo a  <EM ID="hub-77558-14000" CATEG="PESSOA">Zeca Afonso</EM>, Filhos da
<EM ID="hub-77558-14001" CATEG="PESSOA" TIPOREL="relacao_familiar" COREL="hub-77558-14000"> Madrugada</EM>
```

Figura 8: Relação Familiar marcada de maneira equivocada pela primeira versão da Cascata de Transdutores

avaliador utilizado, que precisou ser substituído pela versão disponibilizada pela Linguateca[1]. O mesmo adicionou novas opções de parâmetros de avaliação, restringindo para considerar apenas as relações ou etiquetagens estudadas.

4.2 Relações Sede de e Ocorre em

A segunda categoria de relação explorada foi a *Sede de*. Essas são relações entre uma organização (**ORGANIZACAO**), ou um evento (**ACONTECIMENTO**) em específico, que ocorreu em um certo **LOCAL**. Assim, exploramos as relações que tinham essas três categorias como envolvidas. Com isso, chegamos em algumas regras como: Uma palavra reconhecida como ACONTECIMENTO, seguido de alguma flexão dos verbos acontecer, ocorrer ou realizar que, por sua vez, é acompanhado de uma palavra reconhecida como LOCAL pela Cascata de Transdutores. Por exemplo, "A Topcom acontece anualmente na UFES de Goiabeiras, em Vitória", onde o evento Topcom, que se trata de um evento de programação para graduandos, ocorre na localidade da UFES de Goiabeiras, na cidade de Vitória.

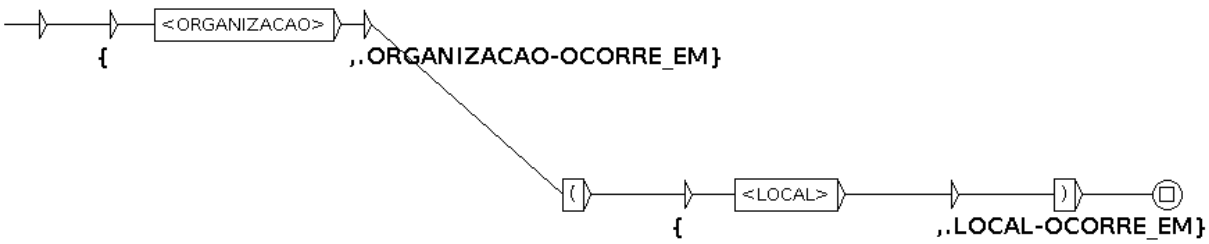

Figura 9: Grafo relacionado a relação *ocorre_em*

A relação *ocorre em* possui uma estrutura semelhante com a que encontramos na explicação acima. Portanto, tratamos o estudo dessas relações junto com a anterior. Com isso, temos a vantagem de estruturar os grafos de forma conjunta. Para tais categorias, encontramos 31 relações exemplos do que denominamos como relações diretas.

[1] https://www.linguateca.pt/harem/avaliacao/?tipo=rerelem

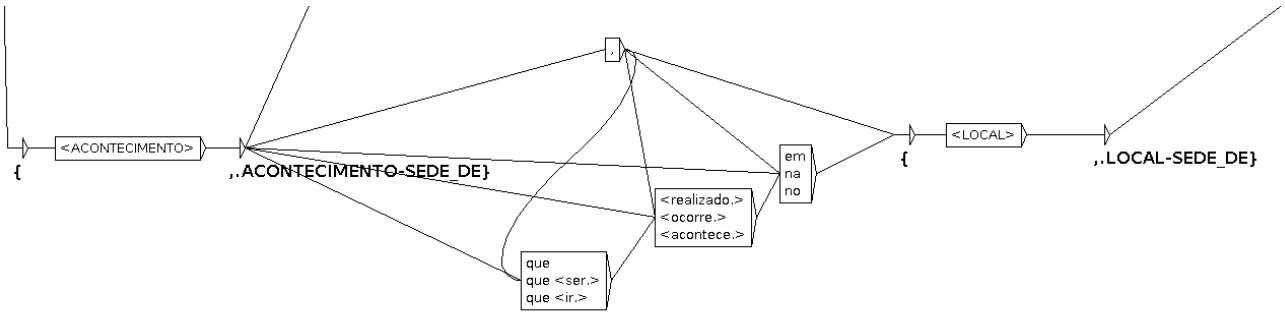

Figura 10: Grafo relacionado a relação *sede_em* entre um Acontecimento e um Local

Porém, houve problemas com as anotações errôneas da Cascata de Transdutores sobre as entidades nomeadas. Por exemplo, na sentença: "...escolasticismo nos finais da **Idade Média** na Europa..." a palavra Idade Média foi anotada como organização. Assim, o padrão: **ORGANIZACAO na LOCAL(Europa)** foi anotado erroneamente como uma relação de *sede em*, interferindo no resultado final. A mesma situação foi repetida com a sentença "Enraizamento da **Educação Ambiental** no Brasil", onde a palavra Educação Ambiental também foi identificada como organização.

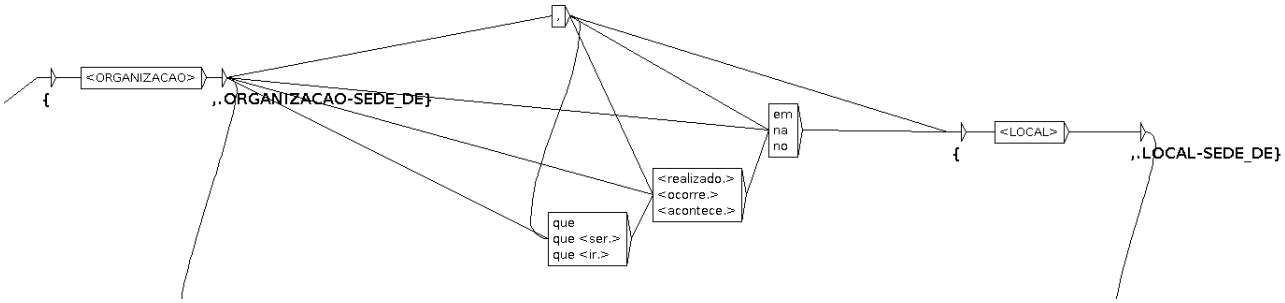

Figura 11: Grafo relacionado a relação *sede_em* entre uma Organização e um Local

O que também aconteceu foi um caso de discordância em relação ao que a anotação fornecida no *corpus* SegundoHarem considera. Na sentença "O acordo ratificado pelo **Conselho da União Europeia** em Lisboa", o nome Conselho da União Europeia é identificado como organização tanto pelos *scripts* quanto pelo resultado anotado, mas a anotação não considera que Lisboa seja a sede deste conselho e, pela sentença, temos a sensação que seja a sede de fato. Por isso, os resultados mais reduzidos dessas relações foi considerado como positivo, já que há discordâncias aceitáveis entre o resultado obtido pela Cascata de Transdutores e a anotação de fato. Ao todo, foram 21 relações identificadas pela Cascata de Transdutores onde, pelo próprio resultado da Tabela 3, cerca de 66% ou 14 relações de forma aproximada, foram anotadas corretamente. Também foi obtido o resultado geral da Cascata, considerando todas as

relações, após a adição dos novos Transdutores. Esses resultados estão na Tabela 4

Corpus	Precisão
SegundoHarem	0.667

Tabela 3: Resultados exclusivos da anotação de relações *ocorre_em* e *sede_de*

Corpus	Precisão
SegundoHarem	0.631

Tabela 4: Resultados gerais após a adição das regras de *ocorre_em* e *sede_de*

4.3 Relações de Inclui e Incluído

Observando tal cenário, foi repetido os passos de: Analisar as relações como um todo, restringir apenas as classificadas como diretas, e em cima dessa lista, criar regras de anotação.

Dessa análise inicial, obtivemos cerca de 511 relações de *inclui* ou *incluido* no nosso *corpus*. Porém, foram poucos cenários de relação direta e, dentro desses, poucas regras bem definidas.

As primeiras ocorrências da relação se davam por período de tempo, onde era citado um século ou um intervalo de tempo e, ao longo do texto, eram citado anos ou "sub-intervalos" dentro do anterior. Tal relação é complexa de ser anotada somente com a nossa ferramenta, pois necessita de cálculos matemáticos para validar o intervalo, além de se tratarem em sua maioria de relações indiretas, que também já explicamos a dificuldade de anotar. Para exemplificar tal cenário, temos as seguintes relações: **<EM ID="H2-dftre765-72" CATEG="TEMPO" > no século XVI</EM>** com as entidades:

1. <EM ID="H2-dftre765-127" CATEG="TEMPO" COREL="H2-dftre765-72" TIPO-REL="incluido" >em 1531</EM>
2. <EM ID="H2-dftre765-105" CATEG="TEMPO" COREL="H2-dftre765-72" TIPO-REL="incluido" >Entre 1540 e 1553</EM>

Outras relações onde notamos uma dificuldade para criar uma regra foram entre grupos. Por exemplo, a seguinte relação **<EM ID="Ytr433-122" CATEG="PESSOA"**

TIPO="INDIVIDUAL" >Ronaldo</EM> volta a treinar com bola no <EM ID="Ytr433-123" CATEG="PESSOA" TIPO="GRUPOMEMBRO" COREL="Ytr433-122" TIPOREL="inclui" >Milan</EM>. Nessa sentença, não há uma regra bem definida relacionando a pessoa Ronaldo ao grupo, no caso o time de futebol italiano, Milan.

Por tais pontos, focamos nas relações entre locais, onde envolve uma cidade com um estado ou um país. Também foram criadas regras para relações entre uma localidade específica, como um ponto turístico e alguma localidade específica em uma cidade. As novas regras serão demonstradas com os exemplos a seguir:

1. <EM ID="hub-63156-24000" CATEG="LOCAL" >Vitória </EM>, cidade do estado do <EM ID="hub-63156-24001" CATEG="LOCAL" TIPOREL="inclui" COREL= "hub-63156-24000" FACS_ORIGEM="LOCAL" FACS_ALVO="LOCAL" >Espírito Santo </EM>, Brasil.

2. nasceu no <EM ID="hub-71248-10" CATEG="LOCAL" >Hospital de São João </EM>, no<EM ID="hub-71248-11" CATEG="LOCAL" TIPOREL="inclui" COREL= "hub-71248-10" FACS_ORIGEM="LOCAL" FACS_ALVO="LOCAL" >Porto</EM>

3. em <EM ID="hub-56266-24000" CATEG="LOCAL" >São Francisco</EM>, nos <EM ID="hub-56266-24001" CATEG="LOCAL" TIPOREL="inclui" COREL="hub-56266-24000" FACS_ORIGEM="LOCAL" FACS_ALVO="LOCAL" >EUA</EM>

Ao todo, foram anotadas 31 relações dessa categoria, onde 15 foram de forma correta. Porém, algumas foram contestadas, como o exemplo: **Viaja com o padrasto, a mãe e os irmãos à <EM ID="aa87333-24000" CATEG="LOCAL" > Ilha Terceira </EM>, nos<EM ID="aa87333-24001" CATEG="LOCAL" TIPOREL="inclui" COREL="aa87333-24000" FACS_ORIGEM="LOCAL" FACS_ALVO="LOCAL" > Açores </EM>, onde vive a família materna.**

Também tivemos o problema de que muitas relações não se tratavam de *inclui* ou *incluido* e sim, de *sede_de* ou enumerações. Mas a ambiguidade dessas relações e os resultados obtidos provaram que essa foi a melhor abordagem, em questão de porcentagem alcançada. A Tabela 5 demonstra os resultados específicos das relações de inclui e incluído no SegundoHaremRelRelem.

Os resultados finais, após a adição de todas as relações na Cascata de Transdutores, foi o apresentado na Tabela 6. A tabela 7 apresenta um resumo de todos os resultados.

Corpus	Precisão
SegundoHarem	0.739

Tabela 5: Resultados exclusivos da anotação de relações *inclui* e *incluído*

Corpus	Precisão
SegundoHarem	0.667

Tabela 6: Resultados gerais finais

A tabela 7 apresenta um resumo de todos os resultados obtidos nessa seção, dividindo os resultados pelos tipos de relações trabalhadas.

Tarefa	Precisão
Relação de Parentesco	0.828
Relação de Ocorre Em	0.667
Relação Incluído	0.739
Resultado Final (Demais tarefas)	0.667

Tabela 7: Resumo das tarefas executadas no trabalho para a criação do reconhecedor de relações

4.4 Resultados em outros corpus

Com os resultados obtidos no *corpus* do SegundoHarem, foi possível notar a evolução na adição de cada regra em específico. Mas agora, com uma versão final da Cascata de Transdutores, foi feito um estudo do desempenho da mesma em outros textos e *corpus*, sem envolver um script avaliador e sim, uma análise manual a partir da saída gerada.

4.4.1 IberLEF

O *corpus* do IberLEF, que obtivemos no trabalho de Collovini et al. (2019), também é um texto anotado com as relações entre entidades. Com a Cascata, foram anotadas dez relações, das quais destacamos cinco:

1. Congresso Internacional de Biodiesel, realizado em Ribeirão Preto. (Relação de "Sede Em")

2. Instituto Internacional de Ecologia, em São Carlos. (Relação de "Sede Em")
3. Universidade de Michigan em Ann Arbor. (Relação de "Sede Em")
4. Instituto Butantan, em São Paulo. (Relação de "Sede Em")
5. Universidade da Califórnia em Berkeley. (Relação de "Sede Em")

Todas as relações que foram etiquetadas nesse *corpus* pela Cascata de Transdutores foram da categoria de "Sede Em". E entre elas, todas foram consideradas assertivas na análise manual.

4.4.2 A Tribuna

Um outro *corpus* utilizado foi "A Tribuna" que foi obtido a partir dos experimentos feitos no trabalho de Magalhaes e Pirovani (2020). Ele se trata de uma compilação de artigos e entrevistas do jornal de mesmo nome, do estado do Espírito Santo. Por se tratar de um *corpus* muito extenso, foi resumido a análise em alguns textos específicos, para ser possível a anotação prévia pelo autor. Por isso, foram levantados quatro textos para serem etiquetados pela Cascata. As etiquetagens que tiveram destaques foram as seguintes:

1. Shopping Praia da Costa, Vila Velha. (Relação de "Sede Em")
2. AT em Família (Relação de "Sede Em")
3. foi morto a tiros na rua da Escadaria São Lucas, na Vila Rubim, em Vitória. (Relação de "Inclui")
4. da Universidade de Liverpool, na Inglaterra (Relação de "Sede Em")

A primeira relação foi considerada assertiva, já que se trata de um *Shopping* e sua região da qual está localizado. Já a segunda foi considerado um erro, porém algo aceitável considerando a margem de acertos, já que tratou a palavra AT como uma organização e pelo contexto, se trata de um tipo de reunião. Por fim, as duas últimas também foram consideradas assertivas já que a relação de inclui da terceira sentença se dá entre a rua da Escadaria e a Vila Rubim. A quarta relação também se dá entre uma universidade e a área de seu campus.

5 ESTUDO DE CASO COM AS RELAÇÕES DE PARENTESCO

Uma das aplicações dos resultados de uma Cascata de Transdutores sobre relações foi feita no trabalho de Magalhaes et al. (2021). Neste trabalho, foi utilizada a Cascata de Transdutores obtida em Magalhaes e Pirovani (2020), em seguida foram inseridas as regras de parentescos, apresentadas na sessão 4.1, utilizando a própria Bíblia, o livro sagrado das religiões cristã e judaica.

As regras foram obtidas principalmente do livro de Mateus, no capítulo 1. Esse texto teve certo destaque, pois é descrito a árvore genealógica de Jesus desde Abraão, que foi aquele do qual Deus prometeu uma longa linhagem no que é chamado de antigo testamento.

A partir dessas regras, foram extraídas as relações e geradas entradas no formato esperado para *Prolog*, uma linguagem voltada para lógica matemática. A escolha se dá porque, através das relações diretas entre pai e filho que a Cascata fornece é possível criar funções da qual: Se A é pai de B e B é pai de C, logo A é avô de C. Assim, é possível trabalhar posteriormente com as inferências lógicas.

No trabalho de Oliveira et al. (2021) foi implementado um *Chatbot* utilizando os resultados dessa Cascata de Transdutores, desenvolvida em Magalhaes et al. (2021), para conseguir responder perguntas na linguagem natural a partir dessa estrutura feita com Prolog. Por exemplo: "Quem é avô de Judá?" seria convertido para uma entrada esperada em Prolog como **IS_GRANDFATHER(X,"Judá")**. Assim, poderia se retornar a resposta em um formato tratado para o usuário, como uma frase em linguagem natural.

Todas as regras obtidas no estudo do trabalho com a Bíblia, também foram incrementadas neste trabalho.

is_father(abraao,isaac).	
is_father(isaac,jacó).	is_father(jacó,juda).
is_father(juda,perez).	is_mother(tamar,perez).
is_father(juda,zera).	is_mother(tamar,zera).
is_father(perez,esrom).	is_father(esrom,arao).
is_father(arao,aminadabe).	is_father(aminadabe,nasom).
...	
is_father(abiude,eliaquim).	is_father(eliaquim,azor).
is_father(azor,sadoque).	is_father(sadoque,aquim).
is_father(aquim,eliude).	is_father(eliude,eleazar).
is_father(eleazar,mata).	is_father(mata,jacó).
is_father(jacó,josé).	**is_father(josé,jesus)**.
is_father(abraao,davi).	is_father(davi,jesus_cristo).
is_mother(maria,jesus).	

Tabela 8: Exemplo da saída obtida e tratada do *corpus* da Bíblia

6 CONCLUSÃO

Neste trabalho foi feita uma evolução dos trabalhos anteriores que foram desenvolvidos em Magalhaes e Pirovani (2020) e Pirovani (2019), voltada para um tópico não abordado anteriormente, as relações. Com o estudo prévio de trabalhos dessa área e dos *corpus* anotados com esse objetivo foram escolhidas como foco as relações de parentesco, identidade, inclusão e "ocorre em", que estão presentes no HAREM.

Para estudar e avaliar os resultados, foram utilizados os recursos disponibilizados no SegundoHaremRelRelem. A facilidade se dava na diversidade de exemplos, o padrão fácil de anotação e a automatização da avaliação com scripts fornecidos. Porém, devido a essa diversidade, muitas relações eram etiquetadas devido ao contexto de um parágrafo ou do texto por um todo e isso, a ferramenta escolhida possibilitou apenas a elaboração de regras dentro de uma única sentença, o que para a abordagem linguística é suficiente. Portanto, os resultados de precisão possuem um indicativo melhor do desempenho do trabalho, pois este visava apenas as relações que a Cascata encontrou. Já a abrangência e medida-F forneciam uma visão geral do quanto a mesma conseguia encontrar no texto, considerando todas as relações e entidades. Isso pode ser evoluído em trabalhos futuros, envolvendo até mesmo uma abordagem de aprendizado de máquina combinada.

Uma dificuldade encontrada foi a anotação de REN que tínhamos como base dos trabalhos anteriores. Pois algumas palavras que não eram reconhecidas pela Cascata como entidades nomeadas, as regras não conseguiam anotar relações das quais elas faziam parte. Além disso, algumas palavras anotadas erroneamente também prejudicaram as novas regras.

Para trabalhos futuros, sugerimos a aplicação dessa Cascata de Transdutores combinado com alguma interface gráfica ou *Chatbot*, assim como o trabalho de Oliveira et al. (2021) que era voltado para bíblia, que exibiria de forma mais fácil e tratada para um usuário, as relações que foram obtidas através de um *corpus* enviado como entrada.

Além disso, como foi visto, a Cascata teve muitas restrições devido a abordagem linguística. Logo, combinar a mesma com outras abordagens pode evoluir os resultados que temos atualmente.

Já sobre a abordagem linguística, também sugerimos como um futuro trabalho, uma combinação com outras abordagens. No trabalho de Abreu (2014), diferentes abordagens de aprendizado de máquina foram aplicadas para a tarefa de ER, tais como o Modelo Oculto de Markov, Campos Aleatórios Condicionais, Modelos de Máxima Entropia, K-Vizinhos mais próximos (k-Nearest-Neighbors - KNN), entre outros. A combinação entre uma abordagem linguística com a de máquina pode entregar resultados até melhores do que foi obtido neste.

REFERÊNCIAS

ABREU, S. C. de. Extração de relações do domínio de organizações para o português. In: . Pontifícia Universidade Católica do Rio Grande do Sul, 2014. Faculdade de Informáca. Disponível em: <http://tede2.pucrs.br/tede2/handle/tede/5248>. Citado 2 vezes nas páginas 10 e 34.

COLLOVINI, S. et al. IberLEF 2019 Portuguese Named Entity Recognition and Relation Extraction Tasks. In: *IberLEF 2019*. [S.l.: s.n.], 2019. Citado 3 vezes nas páginas 15, 19 e 29.

FRIBURGER, N.; MAUREL, D. Finite-state Transducer Cascades to Extract Named Entities in Texts. In: *Theoretical Computer Science*. New york, USA: Elsevier, 2004. p. 93–104. Citado 2 vezes nas páginas 10 e 15.

GROSS, M. The Construction of Local Grammars. In: *SCHABÈS, Y. (eds.). Finite-state language processing, Language, Speech, and Communication*. Cambridge, Mass.: MIT Press, 1997. p. 329–354. Citado na página 10.

GROSS, M. A Bootstrap Method for Constructing Local Grammars. In: *BOKAN, N. (Ed.). Proceedings of the Symposium on Contemporary Mathematics.* Belgrado, Sérvia: University of Belgrad, 1999. p. 229–250. Citado 2 vezes nas páginas 15 e 16.

LINGUATECA. 2019. Disponível em: <http://www.linguateca.pt>. Acesso em: 09 jul. 2019. Citado na página 13.

MAGALHAES, G. et al. Using named entities for recognizing family relationships. In: SBC. *Anais do IX Symposium on Knowledge Discovery, Mining and Learning*. [S.l.], 2021. p. 24–32. Citado na página 31.

MAGALHAES, G.; PIROVANI, J. Gramática Local e Cascata de Transdutores para Reconhecimento de Entidades Nomeadas em diferentes gêneros textuais. In: *XXX Jornada de IC - UFES*. [S.l.: s.n.], 2020. Citado 14 vezes nas páginas , 10, 11, 12, 17, 18, 19, 20, 21, 22, 24, 30, 31 e 33.

MUNIZ, M.; NUNES, M. das G. V.; LAPORTE, E. G. C. UNITEX-PB, a set of flexible language resources for Brazilian Portuguese. In: *Workshop on Technology on Information and Human Language (TIL)*. São Leopoldo, Brazil: [s.n.], 2005. p. 2059–2068. Citado na página 15.

OLIVEIRA, E. de et al. Using relational inference engine to answer questions. In: *LALA*. [S.l.: s.n.], 2021. p. 68–77. Citado 2 vezes nas páginas 31 e 33.

PAUMIER, S. *Unitex 3.2 User Manual*. [S.l.], 2021. 377 p. Acesso em: 24/06/2021. Disponível em: <https://unitexgramlab.org/releases/3.2/man/Unitex-GramLab-3.2-usermanual-en.pdf>. Citado na página 17.

PIROVANI, J. CRF+LG: Uma Abordagem Híbrida para o Reconhecimento de Entidades Nomeadas em Português. In: *Tese (Doutorado) — Universidade Federal do Espírito Santo*. Espírito Santo, Brasil: [s.n.], 2019. Citado 7 vezes nas páginas 10, 11, 15, 16, 18, 22 e 33.

PIROVANI, J.; OLIVEIRA, E. de. Extração de Nomes de Pessoas em Textos em Português: uma Abordagem Usando Gramáticas Locais. In: *Computer on the Beach 2015*. Florianópolis, SC: SBC, 2015. p. 1–10. Citado na página 9.

SANTOS, D.; CARDOSO, N. Reconhecimento de Entidades Mencionadas em Português: Documentação e Actas do HAREM, a Primeira Avaliação Conjunta na Área. In: . Linguateca, 2007. p. 413. Disponível em: <http://www.linguateca.pt/aval_conjunta/LivroHAREM/Livro-SantosCardoso2007.pdf>. Citado na página 13.

SOUZA, E. N. P.; CLARO, D. B. Extração de relações utilizando features diferenciadas para português. *Linguamática*, v. 6, n. 2, p. 57–65, Dez. 2014. Disponível em: <https://linguamatica.com/index.php/linguamatica/article/view/v6n2-4>. Citado 2 vezes nas páginas 10 e 19.