INSTITUTO SUPERIOR TÉCNICO
Universidade Técnica de Lisboa

Extracção de Relações entre Entidades

Daniel Tiago de Almeida Santos

Dissertação para obtenção do Grau de Mestre em
Engenharia Informática e de Computadores

Júri

Presidente:	Professora Doutora Maria dos Remédios Vaz Pereira Lopes Cravo
Orientador:	Professor Doutor Nuno João Neves Mamede
Co-Orientadora:	Professora Doutora Maria Luísa Torres Ribeiro Marques da Silva Coheur
Vogais:	Professor Doutor Jorge Manuel Evangelista Baptista

Outubro 2010

Agradecimentos

Gostaria de agradecer ao meu orientador, o Professor Nuno Mamede, à minha co-orientadora, a Professora Luísa Coheur e ao Professor Jorge Baptista toda a disponibilidade e apoio que me deram durante o desenvolvimento da dissertação.

Um agradecimento especial aos meus pais e ao meu irmão por todo o apoio dado ao longo do curso, este trabalho é dedicado a vocês por me terem aturado durante 24 anos!

Quero agradecer à Ivete pela presença e pelo apoio dado. Apesar de teres aparecido numa fase final deste trabalho, foste um grande suporte e uma grande ajuda sempre que foi necessário. Obrigado por me fazeres rir mesmo em alturas de muito stress.

Agradeço também à minha colega e cara amiga Andreia Maurício por toda a ajuda dada, pelas constantes críticas, sugestões e revisões durante as várias fases do documento.

Por último, mas longe de serem os piores, quero agradecer aos meus colegas e grandes amigos André Negrão, João Paiva, Diogo Paulo, António Gorgulho, David Moreira, Bernardo Gouveia, Gonçalo Rosmaninho, Henrique Campos e Pedro Pedrosa por me ajudarem ao longo do curso e também pelos grandes momentos de diversão que me proporcionaram.

Lisboa, Outubro 2010

Daniel Tiago de Almeida Santos

Aos meus pais e ao meu irmão.

Resumo

Actualmente, existe uma necessidade crescente em retirar informação a partir de um texto de forma automática. Existem várias investigações nesta área como a identificação e classificação de entidades mencionadas, detecção de expressões temporais, extracção de relações entre entidades, entre outras, que tentam satisfazer essa necessidade. Os sistemas podem retirar informação a partir da extracção de relações entre entidades, utilizando-a para diversas aplicações.

Nesta dissertação apresenta-se um sistema que tenta maximizar a informação retirada. Definiram-se as directivas com base nos trabalhos já desenvolvidos não só na língua portuguesa, como na língua inglesa. Analisam-se alguns sistemas, as metodologias utilizadas pelos investigadores e os resultados obtidos.

O sistema desenvolvido utiliza uma abordagem baseada em regras, e está integrado na cadeia de processamento do L^2F, no INESC-ID em Lisboa. Os pormenores mais relevantes em cada categoria são detalhados com o objectivo de explicar o racional utilizado para cada relação.

Para avaliar este trabalho usou-se um corpus de avaliação. Este foi anotado por uma linguista que não participou na fase de desenvolvimento do sistema. O objectivo consiste em obter resultados mais credíveis, permitindo assim, uma melhor análise de resultados.

Abstract

Nowadays, there is a growing need to retrieve information from a text automatically. There are several investigations in this area as named entities recognition, detection of temporal expressions, relation extraction between entities, among others, that tries to satisfy this need. The systems can retrieve information from the extraction of relationships between entities, using it for various applications.

This master thesis presents a system that tries to maximize the information retrieved. The directives were defined based on the work already developed not only in the Portuguese language, as in the English language. This dissertation examines some systems, the methodologies used by the researchers and the results.

This system uses a rules-based approach, and is integrated into the processing chain of L^2F, at INESC-ID Lisboa. The most relevant details in each category are explained in order to understand the decisions made during the implementation.

An evaluation corpus was selected and annotated by a linguistic, in order to perform a more independent evaluation, thus allowing a better analysis of the results.

Palavras Chave
Keywords

Palavras Chave

Processamento de Língua Natural

Extracção de Relações

Sistema Baseado em Regras.

Identificação de Entidades

Keywords

Natural Language Processing

Relations Extraction

Rules Based System

Entities Recognition

Índice

Lista de Figuras

Lista de Tabelas

1 Introdução

1.1 Motivação

Uma relação pode ser definida como um operador com um domínio de argumentos. Os laços familiares, a associação a uma instituição ou empresa, a localização de uma pessoa, objecto ou instituição, entre outros, são exemplos de relações. No caso das relações acima enumeradas, cada uma das entidades constitui os argumentos da relação.

O principal objectivo na tarefa de extracção de relações é obter conhecimento semântico que depois será utilizado para melhorar os sistemas em que essa informação é integrada como, por exemplo, sistemas de pergunta-resposta ou de sumarização de texto.

De um modo geral, as relações entre entidades num texto manifestam-se numa palavra ou expressão que liga as entidades ao mesmo tempo que exprime um predicado semântico. Ao se observar o exemplo *"O Pedro é filho do João"*, verifica-se que *"filho"* é o termo que relaciona as entidades *"Pedro"* e *"João"*. Noutros casos, uma mesma palavra ou expressão exprime simultaneamente os termos operador e argumento. Como se verifica no exemplo *"O Pedro é lisboeta"*, aqui *"lisboeta"* exprime não só a relação *NASCIDO-EM* como o seu argumento *"Lisboa"*. Já na seguinte frase, por exemplo, apresenta-se uma relação de localização entre as entidades *INESC_ID Lisboa* e *Lisboa*:

O ***INESC_ID Lisboa*** *situa-se em* ***Lisboa***

A extracção de relações pode também abranger fenómenos mais complexos, em que os termos, geralmente os argumentos, não se encontram linearmente próximos, ou na mesma frase, sendo necessário algum cálculo, (por exemplo, resolução de anáfora, análise sintáctica, etc.) incluindo inferência.

O interesse em extrair relações entre entidades mencionadas (EM) surge naturalmente, depois da detecção e classificação automática de EM. Assim, este trabalho baseia-se nos trabalhos

iniciados por Luís Romão (Romão, 2007), João Loureiro (Loureiro, 2007) e pelos investigadores participantes na tarefa de Reconhecimento de Entidades Mencionadas do Segundo HAREM (Mota and Santos, 2008).

Actualmente, e tendo em conta o progressivo aumento da capacidade computacional nos últimos anos, a possibilidade de desempenhar tarefas automaticamente torna-se cada vez mais importante. A área de Processamento de Língua Natural (PLN) não é diferente e existem várias aplicações, como os sistemas de pergunta-resposta, os sistemas responsáveis pela sumarização de textos, entre outros, que têm tirado partido desse desenvolvimento e do facto de existirem cada vez mais tarefas automáticas, para obterem um melhor desempenho. O reconhecimento de relações entre EM é uma das tarefas de PLN que pode contribuir para melhorar alguns destes sistemas.

1.2 Objectivos

O objectivo desta dissertação é desenvolver um sistema baseado em regras que consiga extrair relações de forma automática a partir de um texto. As relações extraídas podem ser Relações Familiares, Período de Vida, Localização de Pessoas, Localização de Edifícios e Relações Empresariais.

Tal como foi referido anteriormente, sistemas de pergunta-resposta e sistemas de sumarização de texto são exemplos onde a extracção de relações tem um papel importante.

Actualmente, já existem sistemas, tanto para a língua portuguesa como para a língua inglesa, que realizam a tarefa de extracção de relações. No entanto, os sistemas de reconhecimento de relações na língua portuguesa são ainda bastante simples, extraindo pouca informação dos textos analisados. Já os sistemas para a língua inglesa, apesar de mais avançados, não são facilmente adaptáveis ao português, devido às diferenças na estrutura sintáctica das duas línguas, assim como a ausência de corpus anotados.

Assim, pretende-se com a realização deste trabalho resolver alguns dos problemas encontrados nos sistemas produzidos para a língua portuguesa, através do desenvolvimento de um sistema mais completo.

Uma das dificuldades da tarefa de extracção de relações resulta do facto de as línguas

naturais disporem muitas vezes de mais do que uma forma para expressar a mesma relação, como se verifica, por exemplo, na relação de *Parentesco* pai/filho, em que cada um dos termos difere na ordem dos argumentos, embora semanticamente se possa considerar que se trata da mesma relação. Há também casos, como já foi referido, em que certos termos aglutinam o nome da relação com um dos argumentos. Noutro caso ainda, observam-se alternâncias na estrutura sintáctica das expressões, como por exemplo, *"ter sede em"* tem o mesmo significado que *"estar sedeado em"*.

1.3 Estratégia

A tarefa de extracção de relações será efectuada através de técnicas de processamento de língua natural, utilizando a ferramenta XIP[1], inserida na cadeia de processamento do L^2F (Laboratório de Sistemas de Língua Falada). Esta será descrita detalhadamente no Capítulo 3.

A extracção de relações usa uma abordagem por regras que analisam o contexto e a estrutura das entidades para detectar ou inferir relações. Todas as relações são testadas através de um corpus abrangente com textos retirados de jornais mas também de outras fontes da Internet. Desta forma, pretende-se obter formas distintas de escrita que representam a mesma informação.

1.4 Roteiro

A estrutura deste documento é a seguinte: no próximo capítulo, comparam-se os trabalhos realizados nesta área, com especial ênfase nas relações identificadas, juntamente com uma breve descrição dos sistemas e dos resultados obtidos. No capítulo 3, faz-se uma descrição da arquitectura da solução e do sistema onde este trabalho se enquadra. O capítulo 4 explica a implementação das regras, dando maior ênfase à sintaxe utilizada, às opções tomadas e ao âmbito de cada relação. No capítulo 5, descrevem-se os resultados obtidos e uma análise crítica aos mesmos. Finalmente, o último capítulo apresenta as conclusões relativas à realização deste trabalho juntamente com algumas ideias para trabalho futuro.

[1]Xerox Incremental Parsing

2 Trabalho Relacionado

A investigação na detecção e classificação de relações entre diversos tipos de entidades começou a desenvolver-se devido à necessidade de extrair esta informação dos textos. As avaliações conjuntas foram um forte incentivo ao desenvolvimento nesta área, pois permitem que os investigadores tenham acesso a um corpus de avaliação e, em alguns casos, a um corpus de treino. Desta forma, os investigadores podem comparar as suas abordagens, identificando os pontos fortes e fracos dos seus sistemas.

Estas avaliações conjuntas começaram com o *"Message Understanding Conference"* (MUC), onde, a partir do MUC-6 (Grishman and Sundheim, 1996), se definiram as primeiras directivas para a detecção e classificação de EM.

No MUC surge o conceito de enquadramento que relaciona as entidades com certos atributos como, por exemplo, a localização de uma entidade. Em (Grishman and Sundheim, 1996) encontra-se informação mais detalhada sobre estes enquadramentos e os atributos de cada um. Através deste conceito, é possível estabelecer certas relações entre a entidade que está enquadrada e as entidades presentes nos atributos do enquadramento.

Na última edição do MUC, em 1998, encontram-se referências sobre a extracção de relações como a localização de uma empresa ou o local de trabalho de uma pessoa[1]. As directivas do MUC-7 estão presentes em (Chinchor, 1997).

Em 1999, surge uma avaliação conjunta denominada de *"Automatic Content Extraction"* (ACE) cuja primeira edição ocorre em 2002. Desde então, tem decorrido quase todos os anos. A extracção de relações entre EM está presente em todas as edições do ACE e as directivas têm evoluído ao longo dos anos, tentando maximizar a extracção automática de informação a partir de um texto.

[1]Para mais detalhe sobre os procedimentos adoptados ver `http://www.itl.nist.gov/iad/894.02/related_projects/muc/proceedings/ie_task.html`

Segundo (Doddington et al., 2004), o principal objectivo do ACE é promover o desenvolvimento de aplicações que realizam as tarefas pretendidas, em vez de tentar resolver as necessidades que levaram a estas pesquisas. Desta forma, os sistemas participantes que apenas pretendem extrair relações só se preocupam com essa tarefa, abstraindo-se de outras utilizações para estes sistemas como por exemplo, integração em sistemas de pergunta-resposta.

Na anotação do corpus de avaliação do ACE, são identificadas as relações que estão presentes nas directivas da tarefa. Nesta anotação, existe uma diferenciação entre relações que se encontram explícitas no texto e outras onde é necessário avaliar o contexto (implícitas). Em 2004, o corpus de treino continha 300 mil palavras, o corpus de avaliação tinha cerca de 50 mil palavras e as directivas apresentavam 7 categorias com 23 relações diferentes. É de salientar que toda esta investigação foi feita para a língua inglesa e o âmbito deste trabalho é a língua portuguesa. Como tal, procedeu-se a uma análise das directivas presentes nas avaliações internacionais, de forma a compreender se estas podem ou não ser adaptadas ao português.

Para a língua portuguesa, a principal referência à extracção de relações é a *"Avaliação de Reconhecimento de Entidades Mencionadas"* (HAREM). A sua primeira edição (Cardoso and Santos, 2007) decorreu em 2005 e dela resultam apenas directivas para detecção e classificação de EM. Na segunda edição, decorrida em 2008, surge pela primeira vez uma tarefa para detecção de relações entre entidades, mas numa versão muito simples, praticamente experimental. As directivas encontram-se presentes no livro dessa edição (Mota and Santos, 2008).

O presente capítulo está estruturado em três secções: na primeira secção, descrevem-se as principais directivas encontradas, as relações e os vários subtipos, acompanhados de exemplos ilustrativos do âmbito da relação. Referem-se também os trabalhos de investigação e as avaliações conjuntas em que as relações estão referenciadas. Nas restantes secções, faz-se uma breve análise às metodologias utilizadas pelos sistemas que detectam relações, com maior ênfase em sistemas por regras, e indicam-se os resultados alcançados por cada um dos sistemas aqui apresentados.

2.1 Directivas

Apresentam-se nesta secção as principais directivas adoptadas em diversos trabalhos de investigação realizados na área de extracção de relações.

2.1.1 Localização

Este tipo de relação identifica a localização de algo, seja a residência de uma pessoa, a localização de uma empresa ou mesmo a região onde decorre um determinado evento. Esta categoria faz parte das directivas do Segundo HAREM (Mota and Santos, 2008). Um exemplo deste tipo de relação pode ser visto na seguinte frase:

Frase: *O EURO 2004 realizou-se em Portugal.*

Relação: *[Localizacao (EURO 2004, Portugal)]*

No MUC-6 e no MUC-7, há referências à detecção da relação de localização para identificar, por exemplo, a localização da sede de uma empresa.

No ACE, esta relação encontra-se sempre presente nas directivas, existindo inclusive vários subtipos, de forma a torná-la mais específica. Estes vão ser descritos posteriormente neste relatório.

Em (Zhao and Grishman, 2005), não se especifica qualquer subtipo. Os autores apresentam o seguinte exemplo *"a military base in Germany"*, explicando que é detectada uma relação entre a organização *"military base"* e a localização *"Germany"*.

Em (Roth and Yih, 2007), também se refere a existência desta relação, no entanto, é apenas referente a uma relação entre duas localizações. Como se verifica no exemplo *"(New York, US)"*. Neste artigo é feita uma distinção entre esta relação de localização e outras relações normalmente enquadradas como subtipos de localização como, por exemplo, a residência.

De seguida, descrevem-se vários subtipos desta relação.

2.1.1.1 Residência

Como o próprio nome indica, refere a residência de uma pessoa.

Frase: *O João mora na Avenida da Liberdade.*

Relação: *[Residencia(João, Avenida da Liberdade)]*

Este subtipo é inicialmente introduzido no ACE02[2] com o nome de *"Residence"*, continua no ACE03[3], deixando de haver, no entanto, uma referência explícita a esta relação a partir da edição de 2004[4]. As directivas dessas edições são utilizadas por (Culotta and Sorensen, 2004) estando aí incluída o subtipo *"Residência"*.

Em (Roth and Yih, 2007) também se identifica a residência, mas com um nome diferente, *"Live In"*. Os autores apresentam o seguinte exemplo: *"(Bush, US)"*.

No segundo HAREM (Mota and Santos, 2008), esta relação também está presente, no entanto, é classificada como "Outras".

2.1.1.2 Proximidade

Este subtipo está relacionado com a proximidade geográfica entre duas entidades. Observe-se o seguinte exemplo:

Frase: *Espanha encontra-se próxima de Portugal.*

Relação: *[Proximidade(Espanha, Portugal)]*

Este subtipo faz parte das directivas de todas as edições do ACE, logo, também é adoptada por (Culotta and Sorensen, 2004).

2.1.1.3 Localização de Empresas

Em quase todas as avaliações conjuntas e investigações relacionadas faz-se uma distinção entre a residência de uma pessoa e a localização de uma empresa. Em alguns casos, existe ainda uma terceira categoria relativa à localização de entidades que não são empresas nem pessoas, categoria essa que será descrita posteriormente neste documento. Já a relação *Localização de Empresas* está presente no seguinte exemplo:

[2]Para mais informação consultar *"ACE Evaluation plan version 06"* que se encontra em `http://www.itl.nist.gov/iad/mig/tests/ace/2002/doc/`

[3]Informação mais detalhada em *"evaluation plan"* que está no presente no site `http://www.itl.nist.gov/iad/mig/tests/ace/2003/`

[4]Para mais informação ver a *"version (7) of the 2004 ACE evaluation plan"* que se encontra em `http://www.itl.nist.gov/iad/mig/tests/ace/2004/`

Frase: *A sede da HP é em Palo Alto.*

Relação: *[Localizacao_Empresas(HP, Palo Alto)]*

Este subtipo encontra-se em todas as edições do ACE, no entanto, o nome da mesma foi-se alterando: nas primeiras edições é referida como *"based-in"*, mas a partir de 2005 começou a denominar-se *"org-location"*[5].

(Culotta and Sorensen, 2004) e (Roth and Yih, 2007) identificam as localizações das empresas. os primeiros seguem o ACE03, enquanto os segundos a identificam com o nome de *"orgBased_in"*. Já o sistema *"Snowball"* (Agichtein and Gravano, 2000) foi desenvolvido especificamente para detectar esta relação.

2.1.1.4 Situado

Este subtipo é muito semelhante à Residência e à Localização de Empresa, sendo que a principal diferença reside no facto de a localização se referir a entidades/eventos e não a pessoas nem empresas. Um exemplo é:

Frase: *A Torre de Belém está situada em Lisboa.*

Relação: *[Situado(Torre de Belém, Lisboa)]*

O subtipo *Situado* encontra-se presente em todas as edições do ACE, mas só nas primeiras duas é que há explicitamente uma diferenciação entre Residência e este tipo de localização. (Culotta and Sorensen, 2004) também adoptam esta relação.

2.1.1.5 Povo-de

Este subtipo representa a naturalidade de um povo ou de uma parte dos habitantes e faz parte das directivas do segundo HAREM (Mota and Santos, 2008), estando aí incluído na categoria *"Outras"*. Um exemplo desta relação está na seguinte frase:

[5]Para mais informação consultar *"The official evaluation plan for the ACE 2005"* que se encontra presente em `http://www.itl.nist.gov/iad/mig/tests/ace/2005/`

Frase: *O povo timorense resistiu às atrocidades cometidas pelo governo Indonésio.*

Relação: *[Povo_de(povo timorense, Timor)]*

2.1.1.6 Natural-de

Este subtipo tem como objectivo identificar a naturalidade de uma pessoa, podendo ser o país ou mesmo a região onde nasceu. Tal como a relação anterior, também faz parte das directivas do segundo HAREM (Mota and Santos, 2008), onde está incluída na categoria *"Outras"*. Um exemplo da relação *Natural-de* pode ser visto na seguinte frase.

Frase: *O Hugo é natural de Leiria.*

Relação: *[Natural_de(Hugo, Leiria)]*

2.1.1.7 Avaliações conjuntas - Localização

A tabela 2.2 mapeia as relações e as avaliações conjuntas ou trabalhos de investigação onde as relações de localização se encontram presentes. A tabela 2.1 contém a legenda das abreviaturas utilizadas.

Se uma relação está presente em todas as edições de uma avaliação conjunta, ou de outras investigações, é marcada com um "✓". Caso esta relação só apareça em edições específicas, então, indicam-se as edições onde esteve presente.

Tabela 2.1: Legenda das tabelas

Abreviatura	Nome respectivo
CS	Culotta e Sorensen (Culotta and Sorensen, 2004)
ZG	Zhao e Grishman (Zhao and Grishman, 2005)
AG	Agichtein e Gravano (Agichtein and Gravano, 2000)
RY	Roth e Yih (Roth and Yih, 2007)
MUC	Message Understanding Conference
ACE	Automatic Content Extraction
HAREM	Avaliação de Reconhecimento de Entidades Mencionadas
ARE	Anaphora Resolution Exercise (Orasan et al., 2008)

Tabela 2.2: Resumo dos subtipos de Localização

Relação	HAREM	ACE	MUC	CS	ZG	AG	RY	ARE
Localização	✓	-	-	-	✓	-	✓	-
Residência	-	02-03	-	✓	-	-	✓	-
Proximidade	-	✓	-	✓	-	-	-	-
Localização Empresas	-	✓	-	✓	-	✓	✓	-
Situado	-	✓	MUC-7	✓	-	-	-	-
Povo-de	✓	-	-	-	-	-	-	-
Natural-de	✓	-	-	-	-	-	-	-

2.1.2 Relações entre Pessoas

A identificação de relações entre pessoas, sejam elas familiares ou empresariais, permite não só extrair toda a informação relativa a este tipo de relações a partir dos textos, como também estruturar essa mesma informação, ajudando assim os sistemas de sumarização de texto e de pergunta-resposta.

As *Relações entre Pessoas* têm como subtipos, por exemplo, *filiação*, *avô/avó*, *irmão/irmã*, *sócio*, *cônjuge*, *ligação profissional*, entre outros. Alguns dos subtipos, como por exemplo *primo*, são simétricos, pelo que a ordem dos argumentos da relação não é relevante, ao contrário do que acontece com outros subtipos, como se irá observar mais à frente neste capítulo.

Este tipo de relação é importante, por exemplo, nas biografias e outros textos históricos referentes às dinastias dos Reis de Portugal, uma vez que estes tipos de textos são normalmente ricos em relações familiares.

Tal como na categoria *Localização* existem várias subcategorias ou subtipos, sendo que todas elas, excepto a *"Ligação Profissional"*, estão nas directivas do ACE02, do ACE03 e são adoptadas por (Culotta and Sorensen, 2004). A partir do ACE04, todas as subcategorias relativas a relações familiares agruparam-se numa única, que está definida como *"Family"*.

No Segundo HAREM (Mota and Santos, 2008), há referências a relações familiares, no entanto, estas não são discriminadas e estão enquadradas na subcategoria *"Outras"*.

2.1.2.1 Sócio

Esta subcategoria refere uma relação de sócio entre duas pessoas numa vertente empresarial. A ordem dos argumentos é igual à do texto onde é detectada a relação, na medida em que esta relação é simétrica.

Frase: *O Pedro é sócio do Manuel.*

Relação: *[Sócio(Pedro, Manuel)]*

2.1.2.2 Avó/Avô

Este subtipo identifica uma relação entre duas pessoas, em que uma é avô/avó de outra, sendo que o primeiro argumento da relação é o avô/avó e o segundo é o neto/neta.

Frase: *A Joana é neta da Maria.*

Relação: *[Avó(Maria, Joana)]*

2.1.2.3 Filiação

Esta subcategoria representa uma relação de filiação entre duas entidades, o primeiro argumento é o pai ou a mãe enquanto o segundo é o filho ou a filha.

Frase: *D.Dinis filho de D. Afonso III.*

Relação: *[Pai(D. Afonso III, D.Dinis)]*

2.1.2.4 Irmão/Irmã

Este subtipo refere uma relação entre irmãos, independentemente do género, e a ordem dos argumentos mantém-se igual à do texto.

Frase: *O Pedro é irmão do Miguel.*

Relação: *[Irmão(Pedro, Miguel)]*

2.1.2.5 Cônjuge

Esta subcategoria identifica uma relação entre duas pessoas que são cônjuges. A ordem dos argumentos não é relevante, logo permanece igual à ordem encontrada no texto.

Frase: *O Filipe é o marido da Manuela.*

Relação: *[Cônjuge(Filipe, Manuela)]*

2.1.2.6 Ligação Profissional

Esta subcategoria identifica ligações profissionais entre duas pessoas. Esta é a única subcategoria das *"Relações entre Pessoas"* que não faz parte das directivas do ACE nem de (Culotta and Sorensen, 2004), mas sim do Segundo HAREM. Outra especificidade desta relação é o facto de a relação poder ter três argumentos, pois é necessário especificar a função ou cargo que liga as duas pessoas.

Frase: *O Luís é chefe do José.*

Relação: *[Ligação_Profissional(Luís, José, chefe)]*

2.1.2.7 Outra-Familiar

Esta subcategoria abrange todas as relações familiares que não foram referidas anteriormente como, por exemplo, *"primos"* ou *"tios"*. Como esta relação é muito abrangente, a ordem dos argumentos acaba por não ser relevante. Desta forma, a ordem dos argumentos é a mesma da do texto de onde se extraiu a relação. Note-se que, apesar de se perder alguma informação ao adoptar relações genéricas, caso seja necessário no futuro adicionar uma subcategoria, basta especificar a relação, pois o processo de detecção já estará feito.

Frase: *O Joaquim é tio do André.*

Relação: *[Outra-familiar(Joaquim, André, tio)]*

2.1.2.8 Outra-Pessoal

Este subtipo refere uma relação entre duas pessoas que não são da mesma família nem apresentam uma ligação profissional. Tal como a subcategoria anterior, esta também é muito abrangente e mantém a ordem dos argumentos apresentada no texto.

Frase: *O Pedro é amigo do António.*

Relação: *[Outra-pessoal(Pedro, António)]*

2.1.2.9 Avaliações conjuntas - Relações entre Pessoas

Na tabela 2.3 apresenta-se um resumo das várias subcategorias aqui apresentadas, juntamente com as investigações onde estas foram adoptadas.

Tabela 2.3: Resumo dos subtipos de Relações entre Pessoas

Relação	HAREM	ACE	MUC	CS	ZG	AG	RY	ARE
Sócio	-	02-03	-	✓	-	-	-	-
Família	✓	04-08	-	-	-	-	-	-
Avô/Avó	-	02-03	-	✓	-	-	-	-
Filiação	-	02-03	-	✓	-	-	-	-
Irmão/Irmã	-	02-03	-	✓	-	-	-	-
Cônjuge	-	02-03	-	✓	-	-	-	-
Outra-pessoal	-	02-03	-	✓	-	-	-	-
Outra-familiar	✓	-	-	-	-	-	-	-
Outra-pessoal	-	02-03	-	✓	-	-	-	-

2.1.3 Relações Empresariais

Nesta secção descrevem-se as várias relações entre uma pessoa e uma organização ou empresa, tais como *"Empregado"*, *"Fundador"*, etc. Esta relação também abrange ligações entre duas organizações ou empresas como, por exemplo, *"Parceria"*.

2.1.3.1 Empregado

Este subtipo refere uma relação entre uma pessoa e a empresa onde trabalha. Como exemplo, observa-se a seguinte frase:

Frase: *O Jorge trabalha na Microsoft.*

Relação: *[Empregado(Jorge, Microsoft)]*

Esta relação está nas directivas do ACE02, ACE03 e em (Culotta and Sorensen, 2004) com a designação de *"Staff"*. No ACE04, é dividida e renomeada para *"Employment"*, separando os executivos dos restantes empregados da empresa. Existe ainda uma terceira subcategoria para os empregados que não se conseguem categorizar. A partir do ACE05 volta a existir apenas uma categoria.

Em (Zhao and Grishman, 2005) existe a relação *"EMP-ORG"*, que engloba todas as relações referentes ao emprego de uma pessoa. O mesmo se verifica em (Roth and Yih, 2007), sendo que este atribui o nome *"Work-for"* à sua relação. No segundo HAREM (Mota and Santos, 2008) a relação de empregado pertence à categoria *"Outras"*.

2.1.3.2 Parceria

Esta subcategoria representa uma relação de parceria entre duas empresas seja como parceiros de negócio ou subsidiário. De forma a exemplificar, observe-se a frase:

Frase: *Empresa A é subsidiária da Empresa B.*

Relação: *[Parceria(Empresa A, Empresa B)]*

Esta subcategoria está presente nas directivas do ACE02, do ACE03 e em (Culotta and Sorensen, 2004) com duas subcategorias, *"Affiliate"* e *"Partner"*. No ACE04 surge a relação *"Subsidiary"*, substituindo o *"Affiliate"*, enquanto a relação *"Partner"* mantém-se.

2.1.3.3 Afiliação Geopolítica

Esta relação refere uma afiliação entre uma pessoa e uma entidade geopolítica.

Frase: *O Primeiro-Ministro de Portugal reuniu-se com os seus ministros.*

Relação: *[Afiliação_Geo-Política(Primeiro-Ministro, Portugal)]*

A *"Afiliação Geopolítica"* encontra-se em (Zhao and Grishman, 2005), com o nome de *"GPE-AFF"*. De forma a compreender a relação, os autores dão o seguinte exemplo: *"U.S. businessman"*, onde existe a relação entre a EM *"businessman"* e o país *"U.S."*.

2.1.3.4 Fundador

Este subtipo representa a relação entre uma empresa e a pessoa que a criou.

Frase: *Thomas Watson foi o fundador da IBM.*

Relação: *[Fundador(Thomas Watson, IBM)]*

Exceptuando a edição de 2004, esta relação está presente em todas as edições do ACE e em (Culotta and Sorensen, 2004).

2.1.3.5 Gestão

Esta subcategoria identifica uma relação entre uma pessoa que é gestora de uma empresa ou de uma organização.

Frase: *O João é gestor de recursos humanos da EDP.*

Relação: *[Gestao(João, EDP)]*

Esta subcategoria faz parte das directivas do ACE02, ACE03 e também de (Culotta and Sorensen, 2004), adoptando o nome de *"Management"*.

2.1.3.6 Cliente

Esta relação representa a ligação entre uma empresa e uma pessoa que é sua cliente. Como exemplo, observe-se:

Frase: *O Pedro é um cliente habitual da Portugália.*

Relação: *[Cliente(Pedro, Portugália)]*

Pode-se encontrar esta relação nas directivas de (Culotta and Sorensen, 2004), do ACE02 e do ACE03, deixando de estar presente nas edições seguintes desta avaliação conjunta.

2.1.3.7 Membro

Esta subcategoria identifica uma pessoa que é membro de uma organização.

Frase: *O Filipe é membro da AMI.*

Relação: *[Membro(Filipe, AMI)]*

Esta relação está presente no ACE02, no ACE03 e em (Culotta and Sorensen, 2004) com o nome *"Member"*. No ACE04 não faz parte das directivas, sendo novamente incluída a partir do ACE05, com o nome de *"Membership"*.

2.1.3.8 Proprietário

Este subtipo representa a ligação entre uma empresa e uma pessoa que é sua proprietária.

Frase: *A Joana é proprietária do Restaurante Odivelas.*

Relação: *[Proprietario(Joana, Restaurante Odivelas)]*

Tal como a relação anterior, a relação "Proprietário" está presente com o nome *"owner"* em (Culotta and Sorensen, 2004) e em todas as edições do ACE, excepto na edição de 2004.

2.1.3.9 Accionista

Esta subcategoria identifica a ligação entre uma empresa e uma pessoa que é sua accionista.

Frase: *O José é accionista da REN.*

Relação: *[Accionista(José, REN)]*

A subcategoria está presente nas directivas do ACE desde a edição de 2005, sendo designada de *"Shareholder"*.

2.1.3.10 Investidor

Esta subcategoria representa a ligação de uma pessoa que é investidora de uma empresa, existindo, no entanto, uma diferenciação relativamente à relação anterior, pois estes investidores podem ou não ser accionistas da empresa. Um exemplo desta relação é:

Frase: *O Manuel investiu na Empresa InformáticaOdivelas.*

Relação: *[Investidor(Manuel, Empresa InformáticaOdivelas)]*

Pode-se encontrar esta relação nas directivas do ACE a partir da edição de 2005.

2.1.3.11 Estudante

Este subtipo identifica a relação entre uma pessoa que é estudante de uma instituição. A partir do ACE05, esta relação está sempre presente nas directivas.

Frase: *O Daniel estuda no IST.*

Relação: *[Estudante(Daniel, IST)]*

2.1.3.12 Avaliações conjuntas - Relações Empresariais

Na tabela 2.4 encontra-se um resumo dos diferentes subtipos das Relações Empresariais.

Tabela 2.4: Resumo dos subtipos de Relações Empresariais

Relação	HAREM	ACE	MUC	CS	ZG	AG	RY	ARE
Empregado	✓	02-04	-	✓	✓	-	✓	-
Parceria	-	02-04	-	✓	-	-	-	-
Afiliação Geo-Política	-	-	-	-	✓	-	-	-
Fundador	-	02-03 e 05-08	-	✓	-	-	-	-
Gestão	-	02-03	-	✓	-	-	-	-
Cliente	-	02-03	-	✓	-	-	-	-
Membro	-	02-03 e 05-08	-	✓	-	-	-	-
Proprietário	-	02-03 e 05-08	-	✓	-	-	-	-
Accionista	-	05-08	-	-	-	-	-	-
Investidor	-	05-08	-	-	-	-	-	-
Estudante	-	05-08	-	-	-	-	-	-

2.1.4 Identidade

A relação de identidade estabelece-se entre duas expressões que se referem à mesma entidade. Na sua forma mais simples, encontramos o caso de aposições de siglas e o seu respectivo desenvolvimento (por exemplo: Instituto Superior Técnico (IST)).

Situações mais complexas envolvem a resolução de anáforas, sendo que anáfora é uma relação de correferência entre duas expressões num texto, uma primeira (antecedente) e uma segunda que àquela se refere. Esta última pode ser representada por um pronome, de modo a evitar a repetição do antecedente, (exemplo: *"O João decidiu vir de carro. De facto, ele já estava atrasado"*), ou por uma expressão cuja interpretação pode por vezes subtilizar conhecimento extralinguístico. (exemplo: *"Cavaco Silva decidiu vir de carro. De facto, o PR já estava atrasado"*).

A resolução de anáfora, porém, já sai do âmbito desta dissertação, estando aliás a ser objecto de outro estudo que está a ser realizado em paralelo (Nobre, 2010).

Relativamente à relação de identidade, esta encontra-se presente nas directivas do Segundo HAREM (Mota and Santos, 2008). No MUC-6 e MUC-7 há referências (Ng and Cardie, 2002) à resolução de anáforas, sendo possível a partir desse ponto extrair a relação de identidade.

2.1.4.1 Avaliações conjuntas - Identidade

Na tabela 2.5 encontra-se um resumo das investigações onde a relação de Identidade é extraída.

Tabela 2.5: Resumo da relação de Identidade

Relação	HAREM	ACE	MUC	CS	ZG	AG	RY	ARE
Identidade	✓	-	MUC-6 e MUC-7	-	-	-	-	✓

2.1.5 Outros

Nesta secção agrupam-se algumas relações de natureza mais diversa.

2.1.5.1 Artefacto

Esta relação inclui a ligação entre uma pessoa e um artefacto abrangendo as relações tanto de *Posse* como de *Criação*. A relação de Artefacto está presente em (Zhao and Grishman, 2005).

No ACE02 e no ACE03 existe apenas a relação *"Owned"*, ou seja, de posse. Também se pode encontrar esta relação em (Culotta and Sorensen, 2004). Na edição do ACE04, o *"Artefacto"* tem três subcategorias, *"Utilizador"*, *"Inventor"* e *"Outro"*. Nas edições seguintes, estas subcategorias são eliminadas, ficando todas abrangidas pela categoria Artefacto.

No Segundo HAREM (Mota and Santos, 2008), também existem relações relativas a artefactos, nomeadamente *"Obra de"* e *"Proprietário"*, pertencendo ambas à categoria *"Outras"*.

Frase: *Alfred Nobel inventou a dinamite.*

Relação: *[Obra_de(Alfred Nobel, dinamite)]*

Tabela 2.6: Resumo dos subtipos de Artefacto

Relação	HAREM	ACE	MUC	CS	ZG	AG	RY	ARE
Artefacto	-	05-08	-	-	✓	-	-	-
Posse	✓	02-03	-	✓	-	-	-	-
Utilizador	-	04	-	-	-	-	-	-
Inventor	✓	04	-	-	-	-	-	-
Outros	-	04	-	-	-	-	-	-

2.1.5.2 Outras Afiliações

No ACE é possível encontrar outro tipo de afiliações, cujo âmbito difere das que foram referidas anteriormente.

No ACE04 existe a relação *"Outras-afiliações"*, que tem como subtipos *"Etnia"*, *"Ideologia"* e *"Outros"*. Nas edições seguintes, a estes subtipos juntam-se também a *"Religião"* e a *"Residência"*, dando um novo nome à categoria: *"Afiliação Global"*.

Acrescentou-se também às directivas uma nova subcategoria chamada *"Afiliação Desportiva"*, pertencendo, no entanto, a uma categoria diferente, pois está agrupada no âmbito das *"Afiliações Empresarias"*.

Frase: *Pepe é jogador do Real Madrid.*

Relação: *[Afiliação_Desportiva(Pepe, Real Madrid)]*

Em (Zhao and Grishman, 2005) também há referências a outras afiliações, onde é dado o exemplo de *"Cuban-American people"* que remete para as pessoas cubanas residentes nos EUA.

Na tabela 2.7 está um resumo da relação Artefacto.

Tabela 2.7: Resumo dos subtipos de Outras Afiliações

Relação	HAREM	ACE	MUC	CS	ZG	AG	RY	ARE
Etnia	-	04	-	-	-	-	-	-
Ideologia	-	04	-	-	-	-	-	-
Outros	-	04	-	-	-	-	-	-
Afiliação Global	-	05-08	-	-	✓	-	-	-
Afiliação Desportiva	-	05-08	-	-	-	-	-	-

2.1.5.3 Inclusão

Esta relação representa as relações entre duas entidades, sendo que há uma que está incluída ou faz parte da outra. Um exemplo dado no segundo HAREM (Mota and Santos, 2008) permite uma melhor compreensão desta relação.

Frase: *O deputado social-democrata Fernando Pereira anunciou ontem a sua candidatura à presidência da Comissão Política Distrital de Vila Real do PSD. (...) Ao contrário do que seria legítimo pensar, a candidatura de Fernando Pereira não aparece como resposta aos maus resultados obtidos pelo PSD nas eleições autárquicas.*

Relação: *[Inclusão(Comissão Política Distrital de Vila Real do PSD, PSD)]*

Considera-se que a entidade mencionada na 1ª frase (*"Comissão Política Distrital de Vila Real do PSD"*) está incluída na segunda ocorrência de *"PSD"*.

(Orasan et al., 2008) não têm referências directas a uma inclusão, mas a dois conceitos semelhantes: generalização e especialização. Neste caso, os autores conseguem detectar as relações através da resolução de anáforas. Uma generalização consiste em dar uma definição mais geral a uma entidade (como no exemplo: *"O futebol é um desporto colectivo."*. Neste caso, desporto

colectivo é uma generalização de futebol. A especialização é o oposto, ou seja, é algo que dá uma definição mais concreta, (como na frase: *"Um exemplo de mamífero é a baleia."*). Aqui, a palavra *"baleia"* pretende dar uma definição mais específica de *"mamífero"*.

Tabela 2.8: Resumo da relação de Inclusão

Relação	HAREM	ACE	MUC	CS	ZG	AG	RY	ARE
Inclusão	✓	-	-	-	-	-	-	-
Generalização	-	-	-	-	-	-	-	✓
Especialização	-	-	-	-	-	-	-	✓

2.1.5.4 PER-SOC

Esta relação, que se encontra em (Zhao and Grishman, 2005), pretende identificar uma pessoa que está a falar em nome de outra pessoa. O exemplo dado neste artigo é *"a spokesman for the senator"*, onde está representada uma ligação entre um senador e o seu porta-voz.

Trata-se de uma relação particularmente dependente de um género textual ou situação comunicativa, pelo que carece de generalidade significante para ser integrada num sistema genérico de extracção de relações.

2.1.5.5 Participante

Esta relação representa a relação de participação entre uma entidade e um evento, que pode ser político, desportivo, cultural, religioso, entre outros. Naturalmente, pressupõe a identificação de eventos enquanto entidades mencionadas. Faz parte das directivas do Segundo HAREM (Mota and Santos, 2008).

Frase: *Nélson Évora participou nos Jogos Olímpicos de 2008*

Relação: *[Participante(Nélson Évora, Jogos Olímpicos 2008)]*

2.1.5.6 Personagem de

Esta relação identifica personagens de obras ficcionais (livros, filmes, peças de teatro, entre outros). Tal como a anterior, também faz parte das directivas do Segundo HAREM (Mota and Santos, 2008) e pressupõe a identificação destas obras como entidades mencionadas.

Frase: *Robert Langdon é uma personagem do Código da Vinci.*

Relação: *[Personagem_de(Robert Langdon, Código da Vinci)]*

2.1.5.7 Período de Vida

Este subtipo tem como objectivo identificar o período de vida de uma determinada pessoa e também faz parte das directivas do Segundo HAREM (Mota and Santos, 2008). Um exemplo desta relação é:

Frase: *D. Dinis viveu entre 1261-1325.*

Relação: *[Período_de_vida(D. Dinis, 1261, 1325)]*

2.1.5.8 Representado por

Trata-se da relação que se estabelece entre a EM que designa uma personagem e a EM que designa o actor que representa essa personagem num filme ou peça teatral. Faz parte das directivas do Segundo HAREM (Mota and Santos, 2008) pertencendo à categoria "Outras".

Frase: *No filme Código Da Vinci Robert Langdon é representado por Tom Hanks.*

Relação: *[Representado_por(Robert Langdon, Tom Hanks)]*

2.1.5.9 Praticado em

Esta relação estabelece uma ligação entre um evento e um local. Tal como as anteriores, também está presente no Segundo HAREM (Mota and Santos, 2008).

Frase: *Os Jogos Olímpicos de 2008 decorreram em Pequim*

Relação: *[Praticado_em(Jogos Olímpicos 2008, Pequim)]*

2.1.5.10 Assassinar

Em (Roth and Yih, 2007) existe a relação *"kill"*, ou seja, *"assassinar"*, representando uma ligação entre duas pessoas, sendo que uma delas assassinou a outra. A ordem dos argumentos é relevante, sendo o primeiro argumento o autor do homicídio e o segundo a vítima. Um exemplo demonstrativo é:

> **Frase:** *Oswald foi de acordo com 4 investigações do governos dos EUA o responsável pelo assassinato de John F. Kennedy.*
>
> **Relação:** *[Assassinar(Oswald, John F. Kennedy)]*

Trata-se, naturalmente, de um tipo de relação muito específico que só se justifica em sistemas para domínios específicos. Como as últimas relações são muito específicas, optou-se por agrupá-las num resumo mais genérico:

2.1.5.11 Avaliações conjuntas - Outras relações

Na tabela 2.9 apresenta-se um resumo das relações descritas anteriormente.

Tabela 2.9: Resumo de outras relações

Relação	HAREM	ACE	MUC	CS	ZG	AG	RY	ARE
PER-SOC	-	-	-	-	✓	-	-	-
Participante	✓	-	-	-	-	-	-	-
Personagem de	✓	-	-	-	-	-	-	-
Período de vida	✓	-	-	-	-	-	-	-
Representado por	✓	-	-	-	-	-	-	-
Praticado Em	✓	-	-	-	-	-	-	-
Assassinar	-	-	-	-	-	-	✓	-

2.1.6 Regras de Inferência

Nesta secção descrevem-se as regras de inferência que foram definidas para o segundo HAREM e que são referidas em (Freitas et al., 2009). Existem quatro regras de inferência que têm como objectivo maximizar o conjunto de relações anotadas pelos sistemas.

Estas regras não fazem parte do trabalho proposto nesta dissertação, porque as directivas adoptadas neste trabalho são diferentes das directivas do HAREM, não sendo assim possível aplicá-las.

As regras de inferência são:

1. Se A tem uma relação de identidade com B e este tem uma relação de identidade com C, então pode-se inferir que A tem um relação de identidade com C.
2. Se A inclui B e B inclui C então A inclui C.
3. Se A inclui B e B é onde se situa a sede de C, então a sede de C também se situa em A.
4. Se A tem uma relação de identidade com B e B tem qualquer relação com C, então A também tem essa relação com C.

2.1.7 Directivas desta Dissertação

Após a análise de todas as categorias e subcategorias anteriormente apresentadas, seleccionou-se um conjunto das mesmas, que constitui uma natureza mais geral e, portanto, se reveste de maior interesse no âmbito de uma aplicação genérica do sistema de extracção de relações. A definição e descrição dessas relações está presente no Anexo A. De seguida, descrevem-se os critérios utilizados para determinar quais as relações que irão ser o objecto deste estudo.

Nas *Relações Familiares* foram especificados alguns graus de parentesco que não estão referidos nos trabalhos de investigação referenciados. Retirou-se a subcategoria *"Outras"*, com o objectivo de maximizar a especificidade das relações familiares.

Quanto ao *Período de Vida*, optou-se por dividir essa relação em duas, separando a data de nascimento da data da morte. Desta forma, pode-se retirar informação sobre uma das duas datas sem ser necessário descobrir a outra.

Separou-se a categoria de *Localização* em duas, uma para pessoas e outra para edifícios, sendo que a primeira tem como objectivo extrair informação sobre a residência, a naturalidade e a nacionalidade de uma pessoa e a segunda pretende determinar a localização de empresas ou de monumentos.

Nas *Relações Empresariais* agruparam-se algumas subcategorias, para evitar a redundância presente em relações como *"Accionista"* e *"Proprietário"*. Adicionou-se uma subcategoria, *"Profissão"*, que não está presente nos trabalhos de investigação acima referidos. Esta decisão pretende determinar a função de uma pessoa numa empresa, visto que a relação *"Empregado"* apenas relaciona a pessoa com a empresa na qual trabalha.

Às *Relações Empresariais* juntou-se ainda a subcategoria *"Afiliação Global"*, que tem como objectivo detectar todas as afiliações genéricas de uma pessoa a uma organização de natureza não empresarial, tais como desportivas, políticas, religiosas, entre outras.

2.2 Sistemas

Nesta secção apresenta-se um conjunto sistemas que visam a detecção de relações entre EM no âmbito do Segundo HAREM, em primeiro lugar, e noutros contextos em seguida.

Estes sistemas podem ser agrupados em dois grupos, conforme a metodologia escolhida para resolver o problema: sistemas por regras e sistemas de aprendizagem automática.

De seguida, descrevem-se as várias abordagens utilizadas.

2.2.1 Sistemas por Regras

Este tipo de sistemas extrai as relações entre as entidades através de regras previamente definidas. Quando uma frase preenche os pré-requisitos de uma regra, é extraída a relação associada a essa regra.

No Segundo HAREM (Mota and Santos, 2008), os três sistemas participantes no *"Reconhecimento de Relações entre Entidades Mencionadas"* (ReRelEM) utilizam a abordagem por regras.

Apresentam-se agora estes sistemas:

2.2.1.1 REMBRANDT

O sistema **R**econhecimento de **E**ntidades **M**encionadas **B**aseado em **R**elações e **AN**álise **D**etalhada do **T**exto (REMBRANDT) (Cardoso, 2008) tem uma particularidade em comparação com os outros sistemas, isto é, usa a Wikipédia[6] não só como fonte de desambiguação, mas também como forma de aumentar a informação obtida para as entidades.

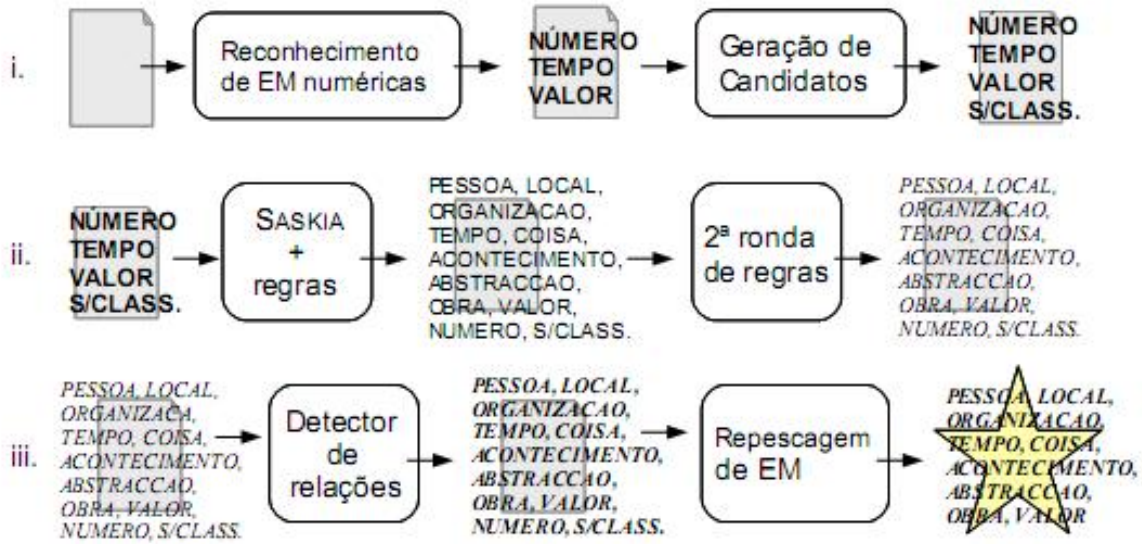

Figura 2.1: Arquitectura do sistema REMBRANDT.

A figura 2.1 retirada de (Cardoso, 2008), descreve a arquitectura do REMBRANDT; os componentes SASKIA e detector de relações são descritos abaixo com mais pormenor.

O SASKIA é uma interface que comunica com a Wikipédia e que verifica se o nome de uma entidade é título de uma página na Wikipédia ou não. Caso seja, é feito um emparelhamento entre ambas, recolhendo as categorias que a Wikipédia atribui à página. Caso não exista essa página, faz-se uma procura nas páginas que tenham o conteúdo o mais semelhante possível. Se essa página existir, então é feito um emparelhamento igual ao parágrafo anterior. Por exemplo, o país Estados Unidos da América tem várias designações: EUA, E.U.A, USA, Estados Unidos, entre outras. Se a página não existir, o SASKIA não devolve nada.

Existem quatro tipos de categorias da Wikipédia reconhecidas pelo SASKIA: categoria normal, autocategoria, categoria de desambiguação e categoria de acrónimo. Esta informação

[6] `http://pt.wikipedia.org/wiki/Pagina_principal`

encontra-se mais detalhada em (Cardoso, 2008). Para além da recolha das categorias, é feita uma classificação das mesmas, com o objectivo de classificar uma EM.

O detector de relações, por seu turno, usa heurísticas simples para detectar as relações existentes. Apresentam-se, de seguida, as heurísticas utilizadas.

Todas as entidades com o mesmo nome, ou que sejam emparelhadas com a mesma página, são classificadas como idênticas. Todas as EM que tenham um nome em comum e em que o nome esteja alinhado a um dos extremos do nome de outra EM são classificadas como idênticas. A relação de *"Identidade"* é então aplicada aos pares de EM classificadas como idênticas.

Para a relação *"Ocorre_em"* é necessário que uma entidade *"Acontecimento"* seja vizinha de outra entidade com categoria *"Local"*. A relação *"Sede_em"* funciona de forma semelhante, mas neste caso a entidade tem a categoria de *"Construção"* em vez de *"Acontecimento"*.

As páginas da Wikipédia que estão emparelhadas com uma entidade são analisadas, de forma a encontrar ligações para páginas de outras entidades, estabelecendo assim uma relação *"Outra"* entre ambas.

O último passo consiste na aplicação de um conjunto de regras gramaticais, para estabelecer relações que ainda não tenham sido detectadas. Essas regras identificam o tipo de relação e o papel de cada EM presente.

2.2.1.2 SEI-Geo

O SEI-Geo (Silveira Chaves, 2008) apenas detecta as relações de *"Inclusão"*, uma vez que está integrado na arquitectura de um sistema de conhecimento geográfico, *Geographic Knowledge Base* (GKB), desenvolvido pelo mesmo autor (Silveira Chaves et al., 2005).

Na figura 2.2, retirado de (Silveira Chaves, 2008), pode-se observar como é feita a detecção de relações. Inicialmente, os textos já segmentados em frases passam pelo componente *"Identificador"*. Este, através de um conjunto de padrões e conceitos de ocorrências de geo-ontologias, determina as frases que têm potencial conteúdo geográfico.

O *"Classificador"* recebe as frases do *"Identificador"* e, consultando as geo-ontologias, faz a desambiguação e detecção de relações.

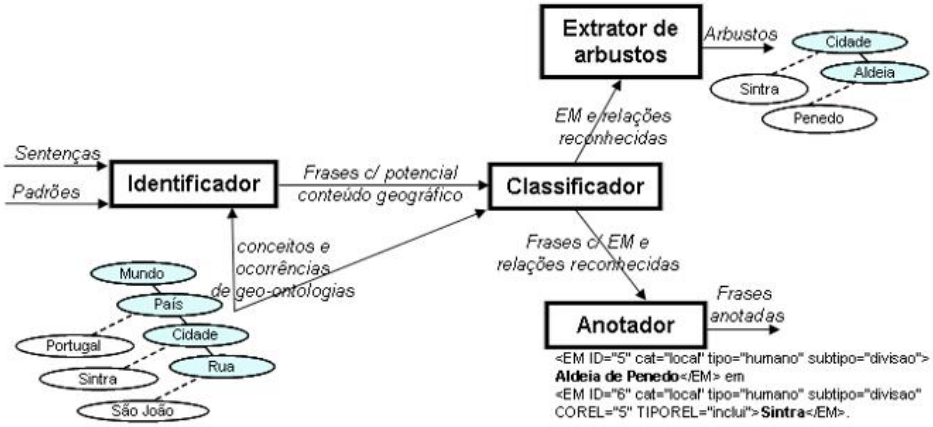

Figura 2.2: Arquitectura do módulo de extracção e anotação de informação geográfica do SEI-Geo .

O componente *"Extrator de Arbustos"* é responsável pela construção de arbustos, sendo cada arbusto composto por, pelo menos, duas entidades e uma relação entre ambas.

Finalmente, o *"Anotador"* faz a anotação no formato desejado.

2.2.1.3 SeRELeP

O SeRELeP (Bruckschen et al., 2008) é um **S**istema de reconhecimento de **RE**lações em textos da **L**íngua **P**ortguesa, que participou no ReRelEM e que se propôs detectar as relações de identidade, ocorrência e inclusão.

Este sistema, tal como os outros participantes do ReRelEM, é baseado em regras. O SeRELeP apenas trata da detecção de relações, pois tanto a identificação como a classificação de EM são feitas por outro sistema, o analisador sintáctico PALAVRAS (Bick, 2000). Na figura 2.3 ilustra-se o processo de anotação de relações.

A ferramenta *"SeRELeP Tools"* recebe a colecção do HAREM e como saída envia o corpus no formato XML[7] para o componente *"SeRELeP"*. A mesma ferramenta envia também o corpus no formato de texto plano para o componente *"Ferramentas Externas"*, que fornece ao *"SeRELeP"* o corpus anotado em XCES[8].

[7] eXtensible Markup Language
[8] XML CES: Corpus Encoding Standard for XML

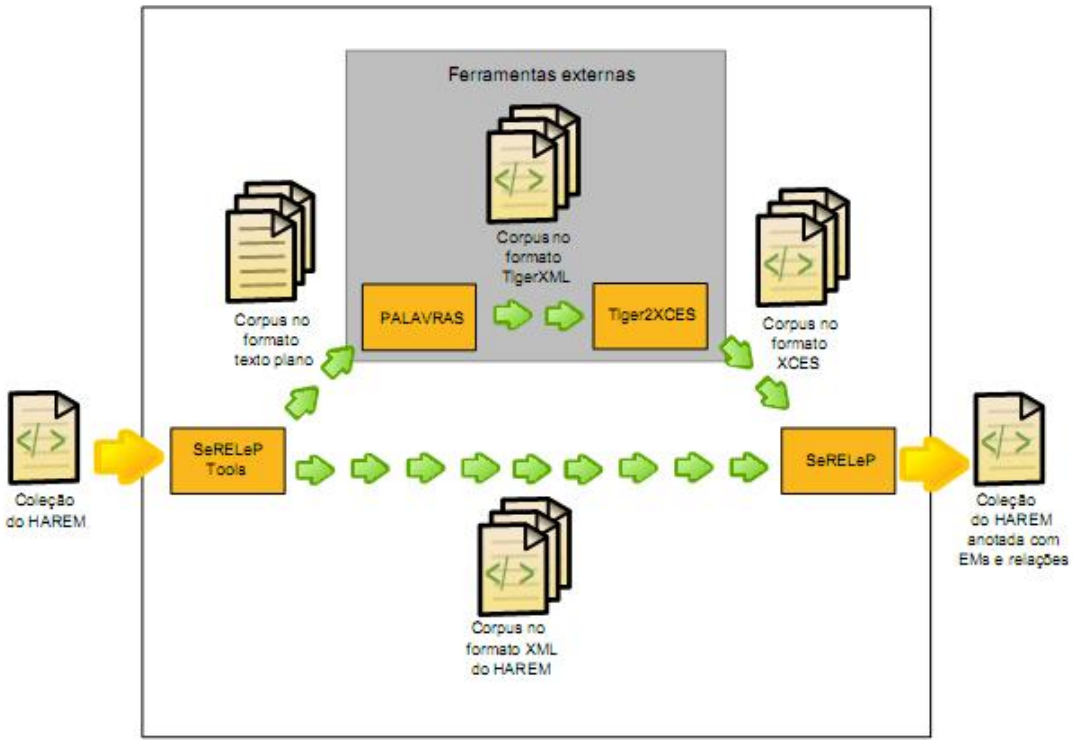

Figura 2.3: Processo de anotação automática de entidades mencionadas e relações retirado de um artigo dos Autores.

O componente *"SeRELeP"*, com base nos dois corpora recebidos, produz uma saída que corresponde à Colecção do HAREM anotada com entidades mencionadas e relações.

Para a detecção de relações, usa-se um conjunto de heurísticas específicas para cada relação, sendo estas executadas de forma sequencial.

Na detecção da relação de *"Identidade"*, verifica-se inicialmente se as EM possuem o mesmo nome. Caso tal não aconteça, analisa-se o nome de uma das entidades verificando se uma é sigla da outra.

Por último, verifica-se se as duas entidades são da categoria *"Pessoa"*. Caso sejam dessa categoria, verifica-se se o nome de uma delas faz parte da expressão que constitui a outra entidade.

Quanto à detecção da relação de *"Inclusão"*, as entidades não podem ter uma relação de identidade entre si. Para além disso, têm de estar presentes na mesma frase e tem de existir uma preposição que indique a presença de uma inclusão, como a preposição *"em"*.

Finalmente, para a detecção da relação *"Ocorre_em"*, analisam-se os casos em que as entida-

des são anotadas como *"Acontecimento"* e *"Local"* ou *"Organização"* e *"Local"*. As regras para esta detecção, são verificadas pela seguinte ordem: primeiro, analisa-se se parte do sintagma da EM *"Acontecimento"* ou *"Organização"* é o nome da entidade *"Local"*; caso não seja, verifica-se a existência de uma entidade *"Local"* na frase onde surgiu a entidade *"Acontecimento"* ou *"Organização"*; por último, faz-se uma procura pela entidade *"Local"* mais próxima da entidade *"Acontecimento"* ou *"Organização"*.

2.2.2 Sistemas de Aprendizagem Automática

Os sistemas de aprendizagem automática usam um corpus de treino para identificar as relações. À medida que se detectam novas relações num texto, estas podem ser acrescentadas à informação já existente.

Existem várias abordagens para desenvolver sistemas de aprendizagem automática, sendo que as duas formas mais comuns são os Sistemas de Aprendizagem Automática Supervisionados e Não Supervisionados.

Considera-se que os sistemas são supervisionados quando os corpora de treino são anotados de forma manual ou de forma automática; já os sistemas não supervisionados não usam corpora anotados.

De seguida, descrevem-se vários sistemas de aprendizagem automática para extracção de relações.

Em (Miller et al., 1998) descreve-se um sistema de aprendizagem automática baseado em modelos probabilísticos com dois módulos principais: o módulo de treino e o módulo de procura.

O módulo de treino recebe anotações sintácticas obtendo informação sobre a estrutura genérica da língua inglesa. Além disso, recebe anotações semânticas, com informações sobre a classificação de EM e sobre as relações a extrair.

Com base nestes dados, o módulo de treino estima os parâmetros de um modelo estatístico. Na presença de uma nova frase, existe um módulo de procura que usa o modelo estatístico, combinando a melhor interpretação sintáctica e semântica.

O sistema *"Snowball"* (Agichtein and Gravano, 2000) consiste num sistema de aprendiza-

gem automática supervisionada que tem um conjunto de dados já anotados, os *"Seed Tuples"* (*"Relações-Semente"*). O sistema pesquisa no texto ocorrências desses mesmos *"Seed Tuples"*.

Assim, se um desses *"Seed Tuples"* indicar *"Sede (Microsoft, Redmont)"*, este sistema procura no texto expressões com essas entidades mencionadas. Caso encontre a expressão *"A sede da Microsoft é em Redmont"*, adiciona a expressão regular *"sede <string1> é em <string2>"*. De seguida, o sistema pesquisa por frases que emparelhem com a expressão regular, obtendo assim novas relações de localização. Sempre que encontrar uma nova relação adiciona aos *"Seed Tuples"*. Finalmente, recomeça o processo com a tabela de dados aumentada, permitindo assim detectar um maior número de relações de localizações das empresas.

(Roth and Yih, 2002) usa o sistema *"SNoW"*, um classificador para múltiplas classes, que foi desenvolvido para tarefas de aprendizagem automática com grandes quantidades de informação. O sistema constrói uma rede de crenças que representam as restrições entre entidades e relações. Para cada frase, usam-se os classificadores existentes, tanto para detecção de entidades, como para detecção de relações. Com base nesses valores e na rede de crenças é calculada qual a categoria mais provável.

(Zelenko et al., 2003) desenvolvem um sistema de aprendizagem automática supervisionada, usando dados classificados automaticamente. A diferença nesta abordagem consiste na utilização dos conceitos de Análise Superficial (*"Shallow Parsing"*) e de Métodos Centrais (*"Kernel methods"*).

O conceito de Análise Superficial, aplicado neste caso às estruturas em árvore, consiste em fazer uma abordagem mais superficial, podendo não dar uma interpretação completa da frase, mas identificando apenas os elementos-chave.

Os Métodos Centrais são funções de correspondência e semelhança entre os nós. As funções de correspondência devolvem 1 caso o tipo e o papel dos nós sejam os mesmos e devolvem 0 no caso contrário. As funções de semelhança devolvem 1, se o texto dos nós for o mesmo, e 0 no caso contrário.

Os autores comparam a sua abordagem com o modelo probabilístico usado por (Miller et al., 1998), distinguindo-a em três aspectos. Primeiro, a análise tradicional é substituída pela Análise Superficial como pré-requisito para extrair relações. Segundo, pretende-se aprender através de modelos locais específicos, contrariamente a um modelo global de texto, dando um

maior ênfase às características locais de uma relação. Finalmente, a última diferença consiste no uso de Métodos Centrais para evitar restrições computacionais ao explorar dependências de longa distância.

São definidos cinco Métodos Centrais Sintácticos, que são definidos através de funções e têm como objectivo quantificar as semelhanças entre duas entidades candidatas a argumentos de uma relação.

Existem dois Métodos Centrais Compostos que utilizam os Métodos Centrais Sintácticos combinando os vários métodos individuais. A Extensão Polinomial (*"Polynomial extension"*) permite obter informação sobre os dois argumentos e sobre o texto de ligação entre eles. O Método Central Completo (*"Full Kernel"*) utiliza todos os Métodos Centrais Sintácticos, em que o valor devolvido por cada um tem o mesmo peso para o resultado final.

Em (Culotta and Sorensen, 2004) define-se um sistema que utiliza um analisador estatístico aplicado a estruturas em árvore. Para cada par de entidades numa frase, é encontrada a subárvore mais pequena, onde ambas as entidades estão incluídas. Assim, é dado um maior ênfase às características locais de cada relação.

(Zhao and Grishman, 2005) usam um sistema de aprendizagem automática supervisionado com a utilização de Métodos Centrais. Estes são divididos em dois tipos: os Métodos Centrais Sintácticos (*"Syntatic Kernels"*) e os Métodos Centrais Compostos (*"Composite Kernels"*).

(Roth and Yih, 2007) também usam um sistema de aprendizagem automática, que utiliza inferência para conseguir derivar novas relações. Para além disso, pretende-se, através da inferência, adicionar restrições ou relações com base nas relações detectadas.

Assim, por exemplo, se A e B representam a mesma localização e se surgir a relação *"João mora em A"*, é necessário inferir a relação *"João mora em B"*. Este tipo de inferências é semelhante às que são adoptadas no Segundo HAREM e que foram descritas anteriormente neste documento.

De seguida, apresenta-se uma tabela (2.10) que esquematiza os vários sistemas referidos e as metodologias utilizadas por cada um deles.

Tabela 2.10: Resumo dos sistemas descritos

Sistema	Regras	Apr. Supervisionada
(Cardoso, 2008)	✓	-
(Silveira Chaves, 2008)	✓	-
(Bruckschen et al., 2008)	✓	-
(Miller et al., 1998)	-	✓
(Agichtein and Gravano, 2000)	-	✓
(Roth and Yih, 2002)	-	✓
(Zelenko et al., 2003)	-	✓
(Culotta and Sorensen, 2004)	-	✓
(Zhao and Grishman, 2005)	-	✓
(Roth and Yih, 2007)	-	✓

2.2.3 Resultados

Nesta secção descrevem-se os resultados obtidos pelos sistemas descritos na secção anterior.

Os sistemas *REMBRANDT*, *SEI-Geo* e *SeRELeP* participaram no segundo HAREM, pelo que é possível fazer uma comparação dos resultados obtidos por estes sistemas. Cada sistema possui várias corridas, ou seja, diversas versões do mesmo sistema definidas de forma a avaliar diferentes condições experimentais. A tabela 2.11 apresenta o melhor resultado obtido por cada um dos sistemas participantes.

Tabela 2.11: Resumo dos resultados obtidos pelos sistemas do Segundo HAREM

Sistema	Precisão	Abrangência	Medida-F
(Cardoso, 2008)	**60%**	35%	**45%**
(Silveira Chaves, 2008)	**90%**	15%	30%
(Bruckschen et al., 2008)	55%	25%	35%

O mecanismo de avaliação utilizado separa a identificação e classificação de entidades das tarefas do ReRelEM, garantindo assim que um sistema não é penalizado várias vezes pelo mesmo erro. São usadas três métricas para avaliar a prestação de cada sistema: a medida-F, a precisão e a abrangência.

Com base na medida-F, a melhor pontuação foi obtida pela corrida 1 do sistema *REMBRANDT*, que obteve 45%, com a precisão e a abrangência a rondar os 60% e 35%, respectivamente. O *SEI-Geo* obteve cerca de 30% de medida-F, com uma precisão superior a 90% e uma abrangência acima dos 15%. O Sistema *SeRELeP* obteve na sua melhor corrida uma medida-F perto dos 35%. A precisão e a abrangência apresentam valores perto dos 55% e 25%

respectivamente.

A tabela 2.12 resume os resultados obtidos pelos sistemas descritos. Nos casos em que os autores testam as relações separadamente apresenta-se o melhor resultado para cada relação.

É importante referir que esta tabela não tem como objectivo comparar os sistemas, visto que todos eles utilizam diferentes directivas e diferentes corpora de avaliação, apenas se reuniu toda esta informação na mesma tabela para facilitar a consulta dos dados.

Tabela 2.12: Resumo dos resultados obtidos pelos sistemas descritos

Sistema	Precisão	Abrangência	Medida-F
(Miller et al., 1998)	81%	64%	71.2%
(Agichtein and Gravano, 2000)	80%	85%	82%
(Roth and Yih, 2002)	-	-	-
"born_in"	87.5%	68.4%	76.6%
"kill"	79.5%	52.8%	62.1%
(Zelenko et al., 2003)	-	-	-
"Person-aff."	91.3%	82.7%	86.8%
"Org-Location"	91.8%	76.3%	83.3%
(Culotta and Sorensen, 2004)	81.2%	51.8%	63.2%
(Zhao and Grishman, 2005)	62.9%	70.5%	70.4%
(Roth and Yih, 2007)	-	-	-
"Located_in"	61.9%	62.9%	59.1%
"Work_for"	79.2%	50.3%	61.4%
"OrgBased_in"	81.7%	50.9%	62.5%
"Live_in"	63.9%	57.3%	59.9%
"Kill"	82.7%	80.8%	81.4%

Os resultados dos restantes sistemas não são directamente comparáveis, pelo que são aqui apresentados a título meramente informativo.

(Miller et al., 1998) testam os seus modelos probabilísticos na detecção de relações, obtendo uma medida-F de 64%.

Em (Agichtein and Gravano, 2000) testa-se o sistema desenvolvido de duas formas: com os sinais de pontuação num texto (*"Snowball"*) e sem a pontuação num texto (*"Snowball Plain"*). Os autores consideravam que a pontuação do texto apenas iria acrescentar ruído na análise das relações. O *"Snowball"* apresenta os melhores resultados com 80% de precisão e 85% de abrangência, o que leva a ter uma medida-F de 82%, contrariando assim a ideia inicial dos autores.

(Roth and Yih, 2002) testam as suas abordagens com duas relações *"kill"* e *"born_in"*. Os

melhores resultados são de 62.1% e 76.6% de medida-F para a primeira e segunda relação, respectivamente.

Em (Zelenko et al., 2003), usam duas relações, *"Person-Affiliation"* e *"Organization-Location"*, para testar as suas abordagens de *"Support Vector Machines"* (SVM). Os melhores resultados são de 86.8% de medida-F para a primeira relação e de 83.3% para a segunda.

Em (Culotta and Sorensen, 2004) usam-se diferentes abordagens para testar o sistema de detecção de relações. O melhor resultado obtido é de 63.2% de medida-F.

Em (Zhao and Grishman, 2005) são avaliadas SVM com diferentes *"kernels"*. O melhor resultado fica pelos 70.4% de medida-F.

(Roth and Yih, 2007) testam o seu sistema lendo frases para identificar entidades e relações, depois testa-se com um conjunto de perguntas e verifica-se se o sistema consegue determinar a resposta. O sistema tem dois classificadores, um para entidades e outro para relações. Os autores testam 5 abordagens diferentes e cada uma é testada para cada relação em separado.

A abordagem omnisciente que consiste em cada classificador assumir que o outro classifica tudo correctamente é a que apresenta melhor medida-F nas relações *"Located_in"* com 59.1%, *"Work_for"* com 61.4%, *"OrgBased_in"* com 62.5% e *"Live_in"* com 59.9%. Para a relação *"Kill"*, o melhor resultado foi obtido pela abordagem dos classificadores separados com uma medida-F de 81.4%. Os autores não apresentaram resultados globais para qualquer abordagem.

3 Arquitectura

Neste capítulo descreve-se a arquitectura geral do sistema em que a tarefa de extracção de relações entre entidades está inserida, isto é, a cadeia de processamento XIP do L^{2}F do INESC-ID[1] em Lisboa.

3.1 Cadeia de Processamento

Esta cadeia de processamento encontra-se ilustrada na figura 3.1 e é baseada em (Mamede, 2007). É usado XML entre os diversos módulos existentes e, sempre que possível, tenta-se transferir a computação do sistema para módulos anteriores ao XIP. De seguida, apresenta-se cada um dos seus módulos.

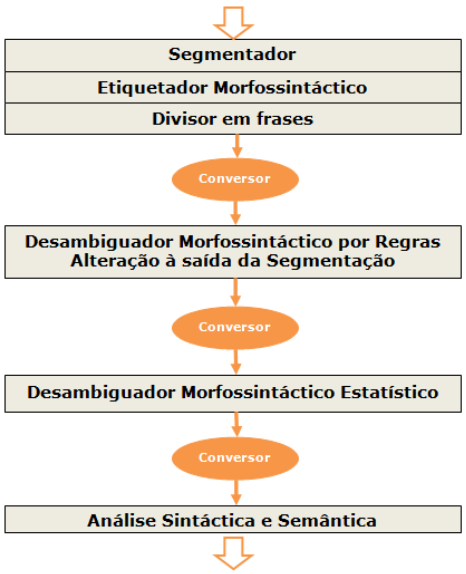

Figura 3.1: Cadeia de Processamento XIP.

O primeiro módulo nesta cadeia de processamento é a *"Segmentador"*, cuja função é dividir

[1]Laboratório de Sistemas de Língua Falada do Instituto de Engenharia de Sistemas e Computadores - Investigação e Desenvolvimento

o texto em unidades léxicas *"Tokens"*. Neste processo são identificados os endereços de e-mail, diferentes tipos de numeração (números ordinais, cardinais, romanos), palavras e sinais de pontuação, entre outros. Ao se considerar a frase "O Pedro foi a Lisboa." como exemplo, verifica-se que a segmentação é "O", "Pedro", "foi", "a", "Lisboa", ".".

O segundo módulo é o *"Etiquetador Morfossintáctico"* que atribui etiquetas morfossintácticas aos segmentos do texto, tarefa realizada através do sistema *Palavroso* (Medeiros, 1995). Este sistema usa o conjunto de etiquetas *"Parole"*, onde se podem encontrar 13 categorias (nome, verbo, adjectivo, pronome, artigo, entre outros). Cada uma das etiquetas tem associada informação relativa à categoria, subcategoria, pessoa, número, género, entre outros. Nesta fase atribuem-se todas as etiquetas possíveis a uma palavra, por exemplo, à palavra *"canto"* é etiquetada com as categorias verbo e nome.

No entanto, existem algumas limitações neste sistema, visto que é um projecto iniciado em 1992. A primeira limitação é o facto de a lematização não ser totalmente adequada à análise sintáctica, pois, por exemplo, os artigos e pronomes apresentam lemas diferentes dependendo do género e do número, quando deveriam ser representados, como lemas, ou seja, pelas formas do masculino e singular. Outros problemas associados a este módulo resultam de os verbos e advérbios não estarem subcategorizados e da dificuldade em introduzir novos traços.

Posteriormente, faz-se a divisão do texto em frases (módulo *"Divisão em Frases"*), utilizando como terminadores os segmentos constituídos por "."; "!"; "?". O resultado é depois convertido para XML, de modo a ser utilizado pelo *RuDriCo* ("Rule Driven Converter") (Pardal, 2007), (Diniz, 2010), um desambiguador morfossintáctico por regras.

Oo sistema *RuDriCo*(*"Desambiguador Morfosintáctico por Regras"*) executa diversas tarefas tais como:

- Fazer uma correcção à saída do Palavroso, alterando os lemas dos pronomes, artigos, advérbios, entre outros (por exemplo, "quaisquer" -> "qualquer");
- Desfazer as palavras contraídas ("nas" é desdobrado em "em" + "as") e na junção de palavras numa única unidade lexical (como por exemplo "Coreia do Norte" que passa a ser tratado como um único *"token"*).
- Fazer a identificação das locuções prepositivas (*"à esquerda de"*) e adverbiais (*"passo a passo"*);

- Tarefa de desambiguação morfossintáctica, escolhendo apenas uma das etiquetas atribuídas pelo sistema *Palavroso*;

A saída do *RuDriCo* é convertida para ser usada pelo desambiguador morfossintáctico *MARV* (*"Desambiguador Morfossintáctico Estatístico"*) (Ribeiro et al., 2003). Este sistema escolhe a etiqueta mais provável para cada palavra que não tenha sido desambiguada pelo *RuDriCo*, utilizando o algoritmo de Viterbi (Jurafsky and Martin, 2000).

Os problemas associados ao *Marv* dizem respeito ao facto de, no cálculo da etiqueta mais provável, utilizar apenas a informação relativa à categoria, subcategoria e frequência lexical, o que em alguns casos é insuficiente. Outro problema reside nos verbos, pois não é escolhido um lema (como exemplo: a palavra "foi" é uma forma do verbo "ir" e "ser"), não sendo possível assim, determinar a estrutura semântica da frase.

O tamanho do corpus de treino é outra limitação do sistema, pois actualmente contém "apenas" 250 mil palavras. Podem existir palavras no corpus de avaliação que não estão presentes no corpus de treino, baixando assim a frequência lexical e, consequentemente, influenciando a etiquetação feita.

De seguida, toda a informação é convertida para o formato aceite pelo *XIP* (Xerox, 2003a), (Xerox, 2003b) e (Xerox, 2003c). que efectua uma divisão em blocos (*"Chunks"*). Depois, é feita a identificação e classificação de entidades mencionadas e calculadas as dependências entre estas entidades.

3.2 Arquitectura XIP

Nesta secção descreve-se a arquitectura do sistema *XIP*.

O sistema *XIP* é um compilador de regras dinâmico com funcionalidades de *"parsing"*, tanto a nível sintáctico como a nível semântico. Uma gramática *XIP* pode ser usada para extrair vários tipos de informação presentes num texto, tais como:

1. Blocos/"Chunks" (sintagmas nominais, sintagmas preposicionais, ...);

2. Dependências;

3. Entidades Mencionadas (locais, pessoas, ...);

4. Papéis semânticos;

5. Intenções comunicativas;

6. Relações de co-referências (anáforas);

Através do sistema *XIP* consegue-se representar diversas características linguísticas, sendo também possível aceder ao contexto circundante. Este sistema é independente da língua, sendo possível criar novas regras sobre as já existentes de forma incremental.

O trabalho referente a esta dissertação enquadra-se na extracção de informação a partir de textos, pelo que será adicionada uma nova funcionalidade ao *XIP* através da introdução de regras, permitindo assim extrair relações não só entre entidades mencionadas, como também entre outras expressões linguísticas. Por exemplo, na frase *"O João vive em Lisboa mas o filho mora em Madrid"*, se o sistema apenas detectar relações entre EM então só identifica a relação de *"Residência"* entre *"João"* e *"Lisboa"*. No entanto, se o sistema detectar relações entre todas as entidades, então são detectadas mais duas relações, uma de *"Residência"* entre *"filho"* e *"Madrid"* e outra de parentesco entre *"João"* e *"filho"*.

As regras aplicam-se após a detecção e classificação de entidades mencionadas. No caso de extracção de relações entre expressões que requeiram a resolução de anáfora, tratando-se de um trabalho que ainda está a ser desenvolvido em paralelo a este, poderá vir a ser pelo menos parcialmente integrado, tendo esta tarefa lugar antes da detecção de relações.

Assim, por exemplo, na presença de uma frase como *"Ele é pai do João."*, apenas é identificada por agora uma relação de parentesco entre *"Ele"* e *"João"*. Futuramente, com a integração da resolução de anáforas, a referência da expressão *"Ele"* será resolvida e estabelecer-se-á a relação de parentesco com a entidade anteriormente referida no texto, isto é, o antecedente de *"Ele"*. O objectivo será maximizar a extracção de informação. As relações que se pretendem identificar e classificar estão descritas no Anexo A.

4 Implementação

Neste capítulo descrevem-se as metodologias e as regras utilizadas para a extracção de relações entre entidades. Para cada categoria descrita nas directivas (ver anexo A) explica-se o fluxo de execução e o raciocínio utilizado para obter as relações.

De forma a extrair correctamente as relações é necessário definir regras com o objectivo de identificar padrões comuns em frases com uma estrutura sintáctica diferente. De seguida, descreve-se a sintaxe dessas regras bem como algumas características do XIP.

Uma regra XIP é composta por 3 partes (Padrão, Condição, Dependências), sendo todas elas opcionais.

```
|Padrão|
  if Condição
Dependências
```

- O *padrão* está relacionado com os nós existentes em cada frase. Um nó é composto por uma ou mais palavras que podem ser nomes, verbos, artigos, preposições, etc. Os nós mais comuns são *SN* (Sintagma Nominal), *SV* (Sintagma Verbal), *SP* (Sintagma Preposicional). Também é possível verificar a presença de certos traços numa palavra como, por exemplo, o género (masculino ou feminino) ou o número (singular ou plural). Os traços mais usados na extracção de relações são o género; o número; o lema de uma palavra; *relative* (que está presente nas palavras que representam uma relação familiar como *"pai"*, *"mãe"*, *"filho"*, *"tio"*, etc); *human* (indica se um determinado substantivo representa ou não uma entidade de tipo *"humano"*); *location* (quando o substantivo é relativo a uma região, cidade ou país); *company* (presente em empresas).

- A *condição* é uma cláusula *"if"* usada para verificar certas condições, como a presença de uma dependência que dá um significado específico à frase, por exemplo, a dependência

SUBJ identifica o sujeito de um verbo. Na frase *"O João vive em Lisboa"*, verifica-se a dependência *SUBJ(vive, João)*. Para extrair as relações, usam-se diferentes dependências. De seguida, explicam-se as mais utilizadas:

- A dependência *SUBJ* identifica o sujeito de um determinado verbo. Por exemplo, na frase: "O João comeu uma maçã.", existe uma relação de *SUBJ* entre o verbo *"comer"* e o sujeito *"João"*.
- A dependência *PREDSUBJ* liga um verbo copulativo como *"ser"* a um nome predicativo, a um adjectivo ou a um advérbio. Por exemplo, na frase *"O João é irmão do Pedro"*, existe a dependência *PREDSUBJ* entre o verbo *"ser"* e o nome *"irmão"*.
- *APPOSIT* relaciona um nome com um aposto. Na seguinte frase: *"O João, irmão do Pedro, mora em Lisboa"* é extraída a seguinte dependência *APPOSIT(João, irmão)*.
- A dependência *COORD* liga elementos presentes numa cadeia de coordenação. Se um verbo está relacionado com um nome e esse nome está relacionado com outro nome, através de uma dependência de coordenação, então a ligação do verbo ao primeiro é estendida também ao segundo nome. Por exemplo, na seguinte frase: *"O João comeu uma maçã e uma laranja"*, o verbo *"comer"* está relacionado não só com *"uma maçã"* mas também com *"uma laranja"*, pois, a par da dependência *CDIR(maçã, comer)* obtém-se também a dependência *CDIR(laranja, comer)*, ou seja, ambos os nomes são considerados complemento directo *(CDIR)* de *"comer"*.
- A dependência *HEAD ("Cabeça")* relaciona o núcleo de um nó com o nó em si. Como exemplo, na frase anterior, o nome *João* é o *HEAD* do nó *"O João"*.
- A última dependência aqui descrita é *MOD* que liga um nó a um complemento desse nó. Estes complementos podem ser sintagmas preposicionais, sintagmas adjectivais, ou orações. Por exemplo na frase: *"O João mora numa casa antiga"* existe uma dependência de *MOD* entre *"casa"* e *"antiga"*.

- Finalmente, a última parte da regra (as *dependências*) determina a acção de uma regra. Neste caso em concreto, cada vez que uma acção é executada extrai-se uma nova relação.

Nas próximas secções descrevem-se os vários tipos de relações extraídas. Indicam-se as dependências e os traços criados para cada tipo de relação, bem como a saída esperada no sistema. Explica-se também o raciocínio efectuado, sendo exemplificado com regras concretas.

4.1 Relações Familiares

O objectivo na extracção de relações familiares consiste em relacionar duas entidades que tenham uma relação familiar entre si. Os textos biográficos são, em norma, muito ricos quanto a este tipo de relações.

Nesta relação pretende-se também especificar a informação obtida, pois dessa forma, são diferenciadas todas as relações familiares através do tipo de relação (Exemplo: Filiação, Tio, Padrinho, Primo, etc).

Criou-se apenas uma dependência, *FAMILY*, que está presente sempre que é extraída uma relação familiar. No entanto, criaram-se vários traços para diferentes graus de parentesco. As relações de *"bisavô"* e *"trisavô"* têm o mesmo traço de *"avô"*, já as expressões *"tio-bisavô"* e *"tio-trisavô"* têm o traço igual a *"tio-avô"*. Apresentam-se de seguida, os traços criados: `uncle, grandparent, cousin, godfather, parent, spouse parent-in-law, grand-uncle, brother-in-law` .

Para além do tipo de relação familiar, também é relevante determinar o género de cada um dos argumentos, permitindo assim reunir certos tipos de relações como, por exemplo, *"pai"* e *"mãe"* dentro do mesmo tipo de relação *PARENT* ("Filiação").

Para isso criaram-se mais 4 traços: `1M, 1F, 2M, 2F`.

Os traços *"1M"* e *"1F"* indicam se o primeiro argumento da relação é do sexo masculino ou feminino, já os traços *"2M"* e *"2F"* indicam o mesmo mas para o segundo argumento.

Assim, na presença da frase *"O Pedro é primo da Filipa"*, o sistema produz a saída:

```
FAMILY_COUSIN_1M_2F(Pedro, Filipa)
```

De seguida, apresentam-se as regras mais importantes, bem como o racional que motivou a sua elaboração.

Uma boa estratégia para detectar relações familiares num texto consiste em procurar padrões relevantes. As palavras que expressam relações familiares são boas pistas a nível lexical. Por exemplo, a palavra *"pai"* pode ser uma boa pista para detectar a presença de uma relação familiar, mas apenas a sua presença não implica necessariamente que a relação exista. Na

frase *"São Pacómio, pai da vida monástica cenobítica"*[1], a palavra *'pai"* é usada no sentido de *"fundador"* e não como relação familiar.

De forma a resolver estes problemas ou outros similares, as regras têm de ser bastante precisas para garantir que uma relação familiar é extraída se e só se os dois argumentos representarem uma entidade *"humana"*; essa entidade pode ser representada por um nome próprio, uma profissão, um título ou mesmo nomes genéricos como *"homem"*, *"mulher"*, *"criança"*, etc.

Após a análise de diversos textos, detectaram-se padrões de frases que indicam a presença de uma relação familiar. Por exemplo, no seguinte padrão de frases:

O João é (pai/tio/primo/irmão) do Pedro

A entidade *"João"* é sempre o sujeito, o verbo *"ser"* relaciona o sujeito com o nome referente ao tipo de relação familiar e o sintagma preposicional que contém *"Pedro"* depende do substantivo que designa a relação. Decidiu-se criar uma única regra que abrangesse este tipo de casos. Só depois se desenvolveram regras que especificam o tipo de relação presente. Tal permite que se crie apenas uma regra para cada padrão de frases em vez de se criarem regras específicas para cada relação, diminuindo assim o número total de regras a implementar.

Apresenta-se agora a regra que permite extrair as relações familiares para os exemplos apresentados anteriormente.

```
  if( HEAD(#2[human],#1) & HEAD(#6[human],#5) &
       PREDSUBJ(#3[lemma:ser],#4[relative]) & PREDSUBJ(#3,#6) &
       SUBJ[PRE](#3,#2) & ~FAMILY(#2,#6) & ~FAMILY(#4,#2,#6) &
       ~MOD[NEG](#3,#7))
FAMILY(#4,#2,#6)
```

Figura 4.1: Regra XIP: Extracção de uma relação familiar

Nesta regra fazem-se as seguintes verificações:

- A primeira linha desta regra tem como objectivo verificar se os dois substantivos presentes na frase se referem a nós com o traço *"humano"*;

[1] `http://hagiaecclesia.blogspot.com/2009/05/sao-pacomio-c.html`

- Verifica-se a existência de uma relação de *PREDSUBJ ("predicativo do sujeito")* entre um verbo que tem o lema *"ser"* e um nó que tem o traço *"relative"*;
- Confere-se a existência da relação *PREDSUBJ* entre o verbo e o nó referente à segunda entidade do tipo *"humano"*;
- Finalmente, verifica-se se o substantivo referente à primeira entidade de tipo *"humano"*, é o sujeito do verbo e se ainda não foi extraída qualquer relação familiar entre aquelas duas entidades;

Caso todas estas condições se verifiquem, cria-se então uma dependência *FAMILY*.

Após a criação da dependência, é necessário especificar a informação que lhe deve estar associada. O primeiro passo consiste em colocar o tipo de relação familiar como um traço da dependência, deixando assim de ser um dos argumentos. A figura 4.2 apresenta a regra que remove o tipo de relação *"primo"* e o coloca com o traço *COUSIN*

```
if( ^FAMILY(#1,#2,#3) & #1[lemma:primo] & #1[masc,sg])
  FAMILY[cousin=+,1M=+](#2,#3)
```

Figura 4.2: Regra XIP: Remoção do tipo de relação familiar dos argumentos

Nesta regra, também é adicionado o traço *1M*, porque se verifica que o substantivo referente ao tipo de relação familiar é masculino e singular. Pode, então, concluir-se que o sujeito da frase de onde se extraiu a relação familiar também é do género masculino.

Para determinar o género do outro argumento da relação criaram-se 4 regras. Na figura 4.3 está uma dessas regras que atribui o traço *1M* caso o primeiro argumento seja do género masculino. Assim, verifica-se se: a primeira entidade da dependência tem o traço masculino e não tem o traço feminino; e se ainda não foi atribuído o traço *1M* à dependência anteriormente.

```
if( ^FAMILY(#1,#2) & ~FAMILY[1M](#1,#2) & ~FAMILY[1F](#1,#2) &
    #1[masc] & ~#1[fem])
      FAMILY[1M=+](#1,#2)
```

Figura 4.3: Regra XIP: Adicionar o traço do género dos argumentos à dependência

Apesar de esta regra identificar a maioria dos casos, existem alguns nomes próprios que são ambíguos, assim como nomes de apelido que não são marcados quanto ao género (Baptista et al., 2006).

Conseguiu-se resolver alguma dessa ambiguidade analisando o género e número do substantivo referente ao tipo de relação familiar, conforme foi explicado na figura 4.2.

A análise do artigo que precede o nome permite resolver alguma ambiguidade, por exemplo, na frase *"o Freitas"*, deduz-se que *"Freitas"* é do sexo masculino devido ao artigo masculino singular que o precede. Para os restantes casos, decidiu-se não incluir o género do argumento.

Certas relações familiares são expressas por expressões idiomáticas, como por exemplo, *"O João e a Maria deram o nó"*. Esta frase significa que *"João"* e *"Maria"* casaram e deve-se extrair a relação *SPOUSE*. Estes casos são diferentes dos anteriores, visto que não se adaptam a todos os tipos de relações familiares, pois são específicos para um certo tipo de relação. Para se conseguirem extrair estas relações, criaram-se regras específicas para cada uma das expressões.

Alguns tipos de relações familiares podem ser acompanhados por outros elementos lexicais, de forma a distingui-los ou para tornar o seu significado mais específico, por exemplo, *"pai adoptivo"*, *"irmão gémeo"*, *"avô materno"*, etc.

Para conseguir extrair estas relações usam-se expressões regulares na indicação dos lemas das palavras. Se o lema de uma palavra for apenas *"pai"*, então a expressão só vai emparelhar com a palavra *"pai"*. Contudo, se o lema passar a ser *"pai(%c*)"* então irá emparelhar não só com *"pai"* mas também com *"pai adoptivo"*. O símbolo *%c* emparelha com qualquer letra, já o símbolo * indica que pode surgir zero ou mais vezes.

Ao contrário de *PARENT* ("pai"), que é uma relação orientada (exemplo: *"João é pai do Pedro"* é diferente de *"Pedro é pai do João"*), há um conjunto de relações familiares, tais como, *"irmão"*, *"primo"*, *"cunhado"*, que são simétricas, isto é, a relação não é orientada (exemplo: *"João é primo do Pedro"* é igual a *"Pedro é primo do João"*), podendo os argumentos serem coordenados (*"Pedro e João são irmãos"*) ou acompanhados de uma cópia pronominal (*"um do outro"*), por exemplo na frase: *"O João e o Pedro são primos"*, o substantivo *"Pedro"* aparece ligado ao sujeito *"João"* através de uma conjunção. Como o tipo de relação *"primo"* é simétrico, pode-se então extrair essa relação. No entanto, se tivermos a frase *"O João e o Pedro são tios"* a expressão em si está sintáctica e semanticamente correcta mas, como tio não é uma relação simétrica, não é possível estabelecer qualquer relação entre as duas entidades.

Até agora, todos exemplos apresentados têm apenas dois argumentos. Outras expressões podem apresentar mais do que duas entidades como, por exemplo *"O João e o Mário são tios*

do Pedro".

Neste exemplo, a relação de *"tio"* estabelece-se entre o *"João"* e o *"Pedro"* e também entre o *"Mário"* e o *"Pedro"*. Para extrair correctamente estes casos, em que duas entidades aparecem coordenadas, implementaram-se as seguintes regras:

```
  if( FAMILY(#1,#2,#3) & #1[pl] & COORD(#4,#2) & COORD(#4,#5) &
     ~FAMILY(#1,#5,#3))
FAMILY(#1,#5,#3)

  if( FAMILY(#1,#2,#3) & COORD(#4,#3) & COORD(#4,#5) &
     ~FAMILY(#1,#2,#5))
FAMILY(#1,#2,#5)
```

Figura 4.4: Regra XIP: Cria uma dependência sempre que se verifica uma coordenação de duas entidades e uma delas tiver uma relação com uma terceira.

Estas regras verificam se existe uma dependência *COORD* entre um dos argumentos de uma relação extraída anteriormente e outra entidade presente na frase. Sempre que esta condição se verifica, propaga-se então essa mesma relação para a nova entidade.

Constitui um caso especial a ausência de sujeito, normalmente porque o mesmo sujeito é referido num momento anterior do discurso, por exemplo na frase anterior. Este tipo de expressões é conhecido por *Anáfora zero*. A resolução de anáfora não faz parte do âmbito deste trabalho, no entanto, a informação presente numa frase deste tipo continua a ser relevante e deve ser extraída. Decidiu-se então criar um nó "fantasma" para os casos em que o sujeito se encontra elidido. Por exemplo, a seguinte frase: *"É tio do Pedro"* é analisada como se fosse *"Ele é tio do Pedro"* permitindo assim extrair uma relação de *"tio"* entre *"Ele"* e *"Pedro"*. Um exemplo de uma regra para este caso está presente na figura 4.5.

```
if( SUBJ[ELIPS](#2,#1) & HEAD(#1,#1) & PREDSUBJ(#2,#3[relative]) &
   (HEAD(#4[people,individual],#5) || HEAD(#4[human],#5) ||
   HEAD(#4[people],#5) || HEAD(#4[profession],#5)) & PREDSUBJ(#2,#4))
     FAMILY(#3,#1,#4)
```

Figura 4.5: Regra XIP: Criação de uma dependência envolvendo um sujeito elíptico.

Se, futuramente, um módulo de resolução de *Anáfora zero* for incorporado no XIP, esta relação ficará com os argumentos correctos.

Quando o substantivo que representa a relação familiar é usado como argumento de outro predicado, também representa o seu sujeito, pelo que é necessário igualmente criar um nó "fantasma". Por exemplo, na frase: *"O João e o tio foram ao cinema"*, *"tio"* exprime não só a relação familiar mas também se refere à entidade que é *"tio do João"*. Para a resolução destas situações, desenvolveu-se uma regra em que se cria um nó virtual preenchido pelo pronome pessoal masculino singular *"Ele"* para representar essa entidade.

```
  if ( (HEAD(#2[people,individual],#1) || HEAD(#2[human],#1) ||
       HEAD(#2[people],#1) || HEAD(#2[profession],#1)) & COORD(#3,#2)
       & COORD(#3,#4[relative]) & ~MOD[POST](#4,#5[human]))
FAMILY(#4,##pron[surface="Ele",lemma="ele",3p=+,sg=+],#2)
```

Figura 4.6: Regra XIP: Criação de um nó virtual para representar a entidade que é assumida pelo nome designativo da relação.

Por último, consideram-se ainda as frases em que se observam pronomes a representar entidades, por exemplo, *"O meu sobrinho Pedro foi ao cinema"*. Neste exemplo, uma das entidades está representada pela palavra *"meu"*, enquanto *"Pedro"* está associado a *"sobrinho"*.

A solução para este tipo de situações consiste novamente em criar um nó para representar a entidade. Para este exemplo em concreto, cria-se um nó *"Eu"*, correspondente à primeira pessoa do singular. (ver figura 4.7).

```
   if ( HEAD(#2[human],#1) & MOD(#4[human],#3[relative])
       & POSS(#2,#5[lemma:meu]))
        FAMILY(#3,#4,##pron[surface="Eu",lemma="eu",1p=+,sg=+]
```

Figura 4.7: Regra XIP: Regra utilizada nos casos de uma entidade representada por um pronome.

4.2 Relação Período de Vida

A relação *"Período de Vida"* tem como objectivo associar uma pessoa à sua data de nascimento ou de óbito. Como nem sempre existe informação sobre ambas as datas, decidiu-se criar relações binárias, associando independentemente uma entidade do tipo *"humano"* a cada uma das datas.

Tal como nas *"Relações Familiares"*, para esta relação criou-se apenas uma dependência *LIFETIME*. De forma a especificar melhor a informação, criaram-se dois traços associados a esta dependência: `born, death`.

Após a criação dos traços e da dependência e perante a frase *"O Pedro nasceu no dia 10 de Fevereiro de 1980"*, a saída do sistema é:

```
LIFETIME_BORN(Pedro, 10 de Fevereiro de 1980.)
```

A estratégia utilizada para a extracção de relações *"Período de Vida"* consiste em procurar frases com expressões temporais, no entanto, nem todas as expressões temporais permitem extrair correctamente a relação. Ao analisar a frase *"O Pedro nasceu em Março"* não é possível tirar grandes conclusões, visto que não é indicado o ano do nascimento.

Para solucionar este problema, decidiu-se que apenas são extraídas relações *"Período de Vida"* se estiver envolvida numa expressão temporal absoluta. Uma expressão temporal absoluta permite situar temporalmente um determinado acontecimento de forma clara. A gramática *XIP* reconhece como expressões de tempo absolutas as expressões com *"Dia, Mês e Ano"*, *"Mês e Ano"* e *"Ano"*.

Tal como na secção anterior, também é necessário garantir que a entidade associada à data é de tipo *"humano"*, evitando assim extrair relações onde uma data está associada a uma empresa ou evento e não a uma pessoa.

De seguida, apresentam-se algumas das regras criadas para extrair este tipo de relações. Utilizando o exemplo dado anteriormente: *"O Pedro nasceu no dia 10 de Fevereiro de 1980"*, a regra implementada para este tipo de relação é a contida na figura 4.8, que faz as seguintes verificações:

- Verifica-se se uma das entidades é de tipo *"humano"*;
- Se essa entidade é sujeito do verbo *"nascer"*;
- Se esse mesmo verbo está ligado a uma expressão temporal do tipo *"data"* e *"absoluta"*, através de uma relação *MOD*;
- Finalmente, verifica-se se não há uma negação na frase.

```
  if( HEAD(#2[human],#1) || & SUBJ[PRE](#3[lemma:nascer],#2) &
      MOD[POST](#3,#4[tipo_tempref:absolut,date])
      & ~MOD[NEG](#3,#5))
LIFETIME[born=+](#2,#4)
```

Figura 4.8: Regra XIP: Regra para extrair uma relação de nascimento.

Caso não houvesse esta última verificação, a relação *LIFETIME* era extraída incorrectamente na seguinte frase: *"O Pedro não nasceu no dia 10 de Fevereiro de 1980"*

Para outros casos é necessário outro tipo de verificações. Por exemplo, nos textos biográficos, é frequente as datas de nascimento e de óbito surgirem logo a seguir ao nome, como na seguinte frase: *"D. Dinis [9 de Outubro de 1261 - 7 de Janeiro de 1325]"*.

Para este tipo de situações fez-se uma análise mais vocacionada para a estrutura sintáctica da frase do que para as dependências extraídas. A regra criada está na figura 4.9. Nesta regra extrai-se tanto a relação de nascimento como a relação de óbito.

```
| NP#1{?*,noun#2[human],PUNCT[left,paren];PUNCT[left,bracket];PUNCT[comma],
 (NP), (PUNCT), NP#3{?*,noun#4[tipo_tempref:absolut,date]}, PUNCT[dash], (NP),
 (PUNCT), NP#5{?*,noun#6[tipo_tempref:absolut,date]}, PUNCT[right,paren];
  PUNCT[right,bracket];PUNCT[comma] |
    if(HEAD(#2,#1))
LIFETIME[born=+](#2,#4),
LIFETIME[death=+](#2,#6)
```

Figura 4.9: Regra XIP: Regra utilizada para extrair datas que estão a preceder o nome.

À semelhança do caso anterior, observam-se outras expressões, por vezes de natureza idiomática, que expressam acções ou eventos referentes a um nascimento ou uma morte, permitindo por isso a extracção da relação *LIFETIME*. Trata-se de expressões como *"ver a luz do dia"*, *"sair da materna sepultura"*, etc.

Implementou-se a regra presente na figura 4.10 para extrair relações em frases como: *"O João terminou com a própria vida no dia 2 de Junho de 1999"*.

O último caso deste tipo de relação está relacionado com a ausência de sujeito. Esse problema foi abordado na secção anterior e a solução mantém-se.

```
| NP#1{?*,noun#2[people,individual];noun#2[human];noun#2[people];
  noun#2[profession]}, VF{verb#3[lemma:terminar]}, PP#4{?*,
  noun#5[lemma:vida]}, PP{?*,noun#6[tipo_tempref:absolut,date]};
  NP{?*,noun#6[tipo_tempref:absolut,date]}|
    if(HEAD(#2,#1) & HEAD(#5,#4) SUBJ[PRE](#3,#2) & MOD[POST](#3,#5))
LIFETIME[death=+](#2,#6)
```

Figura 4.10: Regra XIP: Regra utilizada para a expressão "Terminou com a própria vida".

O sistema cria um nó "fantasma" sempre que uma frase começa com um verbo, esse nó é o pronome *"Ele"* e é usado como argumento da relação. Por exemplo, na seguinte frase: *"Nasceu no dia 3 de Abril de 1988"*, a relação extraída é *LIFETIME_BORN (Ele, 3 de Abril de 1988)*. Posteriormente, com a resolução desta anáfora, o pronome *"Ele"* será *substituído* pela entidade que representa.

4.3 Relação Localização de Pessoas

A relação *"Localização de Pessoas"* associa uma pessoa a um local geográfico. As relações consideradas mais relevantes foram a *"residência"*, a *"naturalidade"*, a *"nacionalidade"* e o *"local de morte"*.

Criou-se então uma dependência que engloba todas estas relações e tem o nome de *PEOPLE-LOCATION*. De forma a especificar a informação, criaram-se quatro traços: `residence, place-of-birth, country-of-birth, place-of-death`.

Com a dependência e os traços criados a saída esperada do sistema para a frase *"O João mora na Avenida da Liberdade"* é:

```
PEOPLE_LOCATION[residence](João, Avenida da Liberdade)
```

A estratégia utilizada para extrair este tipo de relações consiste em fazer pesquisas por palavras como *"mora"*, *"vive"*, *"reside"*, *"nasceu"*, *"é natural de"*, etc. Depois, garante-se que quem reside numa determinada morada é uma entidade do tipo *"humano"* e não uma empresa ou evento. A entidade *"localização"* tanto pode ser uma avenida, um vila, uma cidade, uma região, um país, um continente, etc.

Utilizando o exemplo dado anteriormente, *"O Pedro mora na Avenida da Liberdade."*, a regra usada neste caso está indicado na figura 4.11.

```
  if( SUBJ[PRE](#3[lemma:morar],#2[human]) & HEAD(#2,#1) & HEAD
      (#5[location],#4) & MOD[POST](#3,#5) & ~MOD[NEG](#3,#6))
PEOPLE-LOCATION[residence=+](#2,#5)
```

Figura 4.11: Regra XIP: Regra para extrair uma relação PEOPLE-LOCATION.

De seguida, explicam-se as várias condições que têm de ser satisfeitas para que a relação *PEOPLE-LOCATION* possa ser extraída:

- Verifica-se se o sujeito do verbo *"morar"* é uma entidade do tipo *"humano"*;
- Se o mesmo verbo está associado a uma entidade *"localização"* através de uma relação *MOD*;
- Analisa-se também se a entidade de tipo *"humano"* e a entidade *"localização"* são as cabeças dos nós a que pertencem;
- A última verificação consiste em garantir que não existe uma negação na frase onde a relação está presente;

Em certas expressões é também necessário fazer algumas verificações a nível da estrutura sintáctica da frase, complementando a análise das dependências. Um exemplo disso é a expressão *"A casa do João é na Avenida da República"*. Nesta expressão pode-se substituir "casa" por *"moradia"*, *"vivenda"*, *"apartamento"* ou *"residência"*, etc. Como a estrutura geral não se altera, desenvolveu-se apenas uma regra para todos estes casos. Essa regra está na figura 4.12.

```
| NP{?*,noun#1[lemma:casa]};NP{?*,noun#1[lemma:moradia]}; NP{?*,noun#1
 [lemma:vivenda]};NP{?*,noun#1[lemma:apartamento]}; NP{?*,noun#1[lemma:
 residência]},PP#2{?*,noun#3[human] VF{verb#4[lemma:ser]},
  PP#5{?*,noun#6[location]} |
     if(HEAD(#3,#2) & HEAD(#6,#5) & MOD[POST](#1,#3))
PEOPLE-LOCATION[residence=+](#3,#6)
```

Figura 4.12: Regra XIP: Regra para extrair uma relação Localização de Pessoas com uma estrutura sintáctica específica.

Os exemplos dados até agora referem apenas a residência de uma pessoa. A naturalidade e a nacionalidade constituem casos particulares desta relação.

A naturalidade pode ser extraída quando se encontram expressões como *"é natural de"* ou *"nasceu em"*, entre outras. No caso de *"O João é lisboeta"*, a palavra *"lisboeta"* é um gentílico usado para identificar as pessoas nascidas em Lisboa. Embora, naturalmente, a interpretação de *"residência"* também seja possível, neste caso, dá-se prioridade à interpretação de *"naturalidade"*

De seguida apresenta-se a regra para este tipo de casos.

```
  if(HEAD(#2[human],#1) & ( PREDSUBJ(#3[lemma:ser],#4[gentcity])
   || PREDSUBJ(#3[lemma:ser],#4[gentregion])) & SUBJ[PRE](#3,#2)
    & ~MOD[NEG](#3,#5))
PEOPLE-LOCATION[place-of-birth=+](#2,#4)
```

Figura 4.13: Regra XIP: Regra utilizada para extrair a naturalidade através de gentílicos.

Nesta regra usam-se dois traços que ainda não tinham sido referidos, o *"gentregion"* e o *"gentcity"*, sendo que ambos são atribuídos a gentílicos que representam a naturalidade de uma região ou de uma cidade, respectivamente. Quando as cidades ou regiões são estrangeiras, é então adicionado o traço *"foreign"*.

Em textos biográficos, a par das datas de nascimento ou morte, também se indica muitas vezes os locais associados a estes acontecimentos, por exemplo *"D. Pedro (Coimbra, 8 de Abril de 1320 - Estremoz, 18 de Janeiro de 1367)"*.

Este caso é semelhante ao da relação *LIFETIME*, só que a informação pretendida é o local e não a data de nascimento ou morte. Criou-se então uma regra (ver a figura 4.14) para extrair a naturalidade neste tipo de situações.

```
  | NP#1{?*,noun#2[human]},PUNCT,NP{noun#3[location]},PUNCT,NP{noun
  [tipo_tempref:absolut,date]}, PUNCT,NP{noun[location]},PUNCT,
  NP{noun[tipo_tempref:absolut,date]} |
     if(HEAD(#2,#1))
PEOPLE-LOCATION[place-of-birth=+](#2,#3)
```

Figura 4.14: Regra XIP: Regra utilizada para extrair a naturalidade em textos biográficos.

Para a relação *PLACE-OF-DEATH*, o racional é semelhante ao *PLACE-OF-BIRTH* mas, em vez de expressões como *"é natural de"*, ou *"nasceu em"*, procura-se por expressões como *"morreu em"* ou *"faleceu em"*.

Relativamente à nacionalidade, as expressões são semelhantes às da naturalidade como, por

exemplo, *"O João é português"*; *"O João nasceu em Portugal"*, etc. A diferença consiste nos traços atribuídos à entidade *"localização"*. No caso da relação de naturalidade procuram-se os traços *"location"*, *"gentcity"*, *"gentregion"*, entre outros. No caso da nacionalidade, os traços relevantes são *"country"*, quando a entidade é o nome de um país, ou então *"gentcountry"*, quando a entidade é um gentílico.

A figura 4.15 apresenta a regra para as expressões semelhantes a *"O João é português"*.

```
    if(HEAD(#2[human],#1) & PREDSUBJ(#3[lemma:ser],#4[gentcountry])
     & SUBJ[PRE](#3,#2) & ~MOD[NEG](#3,#5))
PEOPLE-LOCATION[country-of-birth=+](#2,#4)
```

Figura 4.15: Regra XIP: Exemplo de regra utilizada para extrair a nacionalidade.

Tal como nas relações anteriores, a relação *PEOPLE-LOCATION* também tem regras para os casos de omissão de sujeito. Frases como *"Nasceu em Lisboa"*; *"Mora em Lisboa."*, entre outras têm regras para extrair a relação. Como a metodologia é semelhante à descrita nas secções anteriores, decidiu-se não voltar a detalhar este tipo de casos.

4.4 Localização Edifícios

A relação *"Localização Edifícios"* tem como objectivo relacionar uma entidade (empresa, organização ou monumento) com outra entidade que represente uma localização geográfica.

Criou-se uma dependência *BUILDING-LOCATION* que abrange tanto as empresas como os monumentos. Considerou-se que não era necessário distinguir estes dois tipos de entidade.

Ao analisar-se a seguinte frase: *"A Torre de Belém fica em Lisboa."*, a saída esperada é:

```
BUILDING-LOCATION(Torre de Belém, Lisboa)
```

A estratégia utilizada para este tipo de relação é semelhante à anterior e consiste em procurar palavras que exprimam uma relação entre uma empresa ou um monumento e uma localização geográfica como, por exemplo, *"fica em"*, *"situa-se em"*, *"está situada em"*, *"fica situada em"*, *"fica localizada em"*, *"localiza-se em"*, entre outras.

De seguida, apresentam-se alguns exemplos de regras implementadas para extrair este tipo de relações. Na figura 4.16 está a regra desenvolvida para extrair a relação que se verifica no exemplo dado anteriormente: "A Torre de Belém fica em Lisboa".

```
  if( (SUBJ[PRE](#2[lemma:ficar],#1[monument]) || SUBJ[PRE](#2[lemma:situar],
  #1[monument]) ||  SUBJ[PRE](#2[lemma:localizar],#1[monument])) & MOD[POST]
  (#2,#3[location]) & ~MOD[NEG](#2,#4))
BUILDING-LOCATION(#1,#3)
```

Figura 4.16: Regra XIP: Exemplo de regra onde é extraída uma relação BUILDING-LOCATION.

Nesta regras verificam-se certas condições, como o facto de o sujeito do verbo *"ficar"* ser uma entidade com o traço *"monument"*. Naturalmente, tal pressupõe a prévia identificação e classificação dessa entidade. O verbo tem de estar associado a uma entidade *"localização"* através de uma dependência *MOD* e, finalmente, verifica-se se não existe nenhuma negação na frase.

Por outro lado, estas expressões podem sofrer diferentes transformações. Assim, a par da construção passiva com *"ser"*, também se encontra a passiva com *"se"* (*"pode-se visitar"*) e o complemento de lugar pode deslocar-se na frase (*"Em Lisboa pode-se visitar a Torre de Belém."*, *"Pode-se visitar a Torre de Belém em Lisboa."*). Além disso, a presença de um verbo auxiliar (modal), *"poder"*, não altera a construção básica do verbo, sendo capturada através de um tratamento prévio das cadeias verbais com verbos auxiliares.

A regra que extrai todas estas relações está na figura 4.17.

```
  if(HEAD(#1[location],#2) & ( VLINK(#3[lemma:poder],#4[lemma:visitar])
  || VLINK(#3,#4[lemma:ver])) & MOD[POST](#4,#5[monument])
   & ~MOD[NEG](#4,#6))
BUILDING-LOCATION(#5,#1)
```

Figura 4.17: Regra XIP: Exemplo de regra a localização de um monumento.

No caso de localizações de empresas ou organizações, a par dos predicados de localização mais gerais, encontram-se os termos *"sede"*, *"sedeado"*, *"sedear"*. No caso do nome predicativo *"sede"*, este pode surgir como cabeça de um sintagma nominal complexo em frases como *"A sede da Microsoft é em Lisboa"* ou numa frase com *"ter"*, *"A Microsoft tem (a sua) sede em*

Lisboa".

Na relação *BUILDING-LOCATION* também existem casos de ausência de sujeito como, por exemplo, a frase *"Fica situado em Lisboa"*. A solução para estes casos é semelhante ao que foi explicado nas secções anteriores.

O último caso apresentado nesta secção está relacionado com a anáfora. Ao contrário do caso anterior, não existe uma ausência de sujeito mas sim pronome que representa a mesma entidade. Como exemplo, temos *"A sua sede é em Lisboa"*. A relação é extraída e o pronome *"sua"* é um dos argumentos. Quando a resolução de anáforas for integrada no sistema, então o pronome será substituído pela entidade correcta.

4.5 Relação Empresarial

Com esta relação pretende-se associar uma entidade de tipo *"humano"* a uma empresa ou organização. Essa relação pode ser de natureza diversa: *"empregado"*, *"cliente"*, *"fundador"*, entre outras. Se a relação é expressa por um nome de cargo ou profissão, extrai-se então este subtipo de relação empresarial entre a pessoa e o cargo, além da relação com a empresa ou instituição.

A dependência que exprime esta relação é a dependência genérica *BUSINESS*. Vários traços permitem especificar melhor a natureza desta relação: `employee, profession, founder client, owner, affiliation`.

A saída esperada do sistema para a frase: *"O João trabalha na Microsoft."* é:

```
BUSINESS_EMPLOYEE(João, Microsoft)
```

A estratégia utilizada para extrair relações de empregado consiste em pesquisar por palavras como *"ser empregado em"*, *"trabalha em"*, entre outras. Se estiver associado a essa expressão uma entidade referente a uma pessoa e outra entidade referente a uma empresa ou organização, pode, então, extrair-se a relação.

A regra presente na figura 4.18 é a que extrai a relação em frases como, por exemplo, *"O João trabalha na Microsoft"*.

```
if( HEAD(#2[human],#1) & SUBJ[PRE](#3[lemma:trabalhar],#2)
 & ( MOD[POST](#3,#4[org]) ||MOD[POST](#3,#4[company]) &
 ~MOD[NEG](#3,#5)))
BUSINESS[employee=+](#2,#4)
```

Figura 4.18: Regra XIP: Exemplo de regra para uma relação de empregado.

As verificações feitas nesta regra são semelhantes às das secções anteriores. Verifica-se se o sujeito do verbo é uma entidade do tipo *"humano"*; se o lema do verbo é *"trabalhar"* e se esse mesmo verbo tem uma dependência *MOD* com a entidade *"empresa"* ou *"organização"*. Finalmente, garante-se que não existem negações na expressão.

Para extrair relações de profissão utiliza-se um léxico existente no sistema que contém uma lista de profissões e cargos. Depois, verifica-se se existe uma entidade do tipo *"humano"* associada à profissão. Tal pode ser expresso por uma frase predicativa/atributiva, *"O João é carpinteiro"*, ou sob a forma de aposto, no caso de certos nomes de profissão, *"O engenheiro João"*.

Na figura 4.19 encontra-se a regra para o primeiro exemplo dado: *"O João é carpinteiro"*.

```
if( HEAD(#2[human],#1) & SUBJ[PRE](#3[lemma:ser],#2)
& (PREDSUBJ(#3,#4[cargo]) || PREDSUBJ(#3,#4[profession]))
& ~MOD[NEG](#3,#5))
     BUSINESS[profession=+](#2,#4)
```

Figura 4.19: Regra XIP: Exemplo de regra para uma relação de profissão.

Decidiu-se fazer uma distinção entre profissões e cargos. Sempre que o substantivo referente à profissão tem também o traço *"cargo"*, esse é igualmente adicionado à dependência *BUSINESS*; por exemplo, para a frase *"O João é embaixador"*, a saída do sistema é a seguinte:

```
BUSINESS_PROFESSION_CARGO(João, embaixador)
```

Em certas expressões tanto a relação de empregado como a de profissão estão presentes na mesma frase, como, por exemplo, *"O João é engenheiro na Microsoft"*.

Para este tipo de situação utilizam-se regras diferentes para extrair cada uma das relações.

Cada uma dessas regras é específica para a extracção de apenas uma relação, seja *"profissão"* ou *"emprego"*.

Relativamente à relação de *"fundador"* também há palavras-chave associadas como *"fundar"* ou *"criar"*. As regras desenvolvidas para esta relação baseiam-se nestas palavras, nas relações entre as mesmas e entidades humanas ou entidades referentes a empresas ou organizações.

Um exemplo de uma regra está presente na figura 4.20. Esta regra extrai a relação de fundador em frases como: *"O Bill Gates é o fundador da Microsoft"*.

```
  if( HEAD(#2[human],#1) & SUBJ(#3[lemma:ser],#2) & PREDSUBJ(#3,#4
  [lemma:fundador]) & ( PREDSUBJ(#3,#5[company]) || PREDSUBJ(#3,#5[org]))
    & ~MOD[NEG](#3,#6))
BUSINESS[founder=+](#2,#5)
```

Figura 4.20: Regra XIP: Exemplo de regra para uma relação de fundador.

O termo *"pai"* é um caso especial para a relação de fundador, quando se verifica entre uma entidade de tipo *"humano"* e uma entidade de tipo *"empresa"*. Criou-se uma regra especial para dar conta deste caso.

A extracção da relação *"cliente"* está associada a chaves lexicais, como *"comprar em"*, *"cliente"* ou *"ser cliente de"*. A regra para este último caso encontra-se na figura 4.21.

```
  if( HEAD(#2[human],#1) || & SUBJ[PRE](#3[lemma:ser],#2) & PREDSUBJ
  (#3,#4[lemma:cliente]) & (PREDSUBJ(#3,#5[company]) || PREDSUBJ(#3,
  #5[org])) & ~MOD[NEG](#3,#6))
  BUSINESS[client=+](#2,#5)
```

Figura 4.21: Regra XIP: Exemplo de regra para uma relação de cliente.

A relação de *OWNER* abrange não só os donos de empresas como também os seus accionistas. A estratégia utilizada consiste em procurar padrões onde palavras como *"dono"*, *"proprietário"* e *"accionista"* estejam presentes e estejam associadas a entidades do tipo *"humano"* e do tipo *"empresa"*.

De seguida, apresenta-se uma regra utilizada para a seguinte frase: *"O João é accionista da EDP"*.

```
  if( HEAD(#2[human],#1) || & SUBJ[PRE](#3[lemma:ser],#2) & PREDSUBJ(
  #3,#4[lemma:accionista]) & (MOD[POST](#4,#5[company]) || MOD[POST](
  #4,#5[org])) & ~MOD[NEG](#3,#6))
BUSINESS[owner=+](#2,#5)
```

Figura 4.22: Regra XIP: Exemplo de regra para uma relação de proprietário.

A relação *AFFILIATION* pretende abranger tipos de afiliação que podem ser de natureza muito diversa: desportiva, política, religiosa e outras.

À semelhança da relação de *"localização"*, as estratégias utilizadas para a extracção desta relação decorrem dos tipos de construção linguística que as exprimem. Por um lado, padrões envolvendo entidades do tipo *"instituição"* e termos como *"membro"*, *"sócio"*, etc. Por outro lado, os adjectivos de afiliação, tais como *"portista"*, *"socialista"*, *"católico"*, etc.

Na figura 4.23 está um exemplo de uma regra que extrai a relação presente nesta frase: *"O João é membro da UNICEF"*.

```
  if( HEAD(#2[human],#1) & SUBJ[PRE](#3[lemma:ser],#2) & (PREDSUBJ(#3,#4
  [lemma:sócio]) || PREDSUBJ(#3,#4[lemma:membro])) & PREDSUBJ(#3,#5
  [institution]) & ~MOD[NEG](#3,#6))
BUSINESS[global-affiliation=+](#2,#5)
```

Figura 4.23: Regra XIP: Exemplo de regra para uma relação de afiliação.

Em todas as relações presentes nesta secção existem regras para a situação de ausência de sujeito, mas como a metodologia é semelhante às secções anteriores, não se detalha a estratégia utilizada.

4.6 Síntese

As estratégias descritas neste capítulo têm como objectivo abranger os casos mais comuns de cada tipo de relação.

Todas elas consistiram em pesquisar textos reais, como artigos de jornal, anúncios de eventos, biografias online, entre outros, procurando os padrões mais comuns que expressam uma

determinada relação. Depois, para cada padrão de frases encontrado, desenvolveu-se um conjunto de regras responsáveis pela extracção de uma determinada relação.

No desenvolvimento das regras privilegiou-se a precisão das mesmas, com o objectivo de minimizar o número de relações extraídas de forma incorrecta.

Na tabela 4.1 está o número total de regras desenvolvidas para cada categoria de relações.

Tabela 4.1: Número de regras por categoria

Categoria	N.º de Regras
FAMILY	116
LIFETIME	29
PEOPLE-LOCATION	33
BUILDING-LOCATION	16
BUSINESS	74
TOTAL	268

5 Avaliação e Resultados

Neste capítulo descrevem-se os procedimentos utilizados para a avaliação deste trabalho e faz-se uma análise crítica aos resultados obtidos, identificando as falhas ocorridas e explicando os motivos que originam as mesmas.

Na secção 5.1 apresentam-se as métricas de avaliação utilizadas.

A secção 5.2 apresenta os resultados gerais obtidos pelo sistema e também resultados discriminados pelos diferentes tipos de relações extraídas.

Não é possível comparar directamente estes resultados com os resultados obtidos pelos sistemas participantes no HAREM, porque as directivas são diferentes e, contrariamente à avaliação conjunta, este trabalho extrai relações entre qualquer tipo de entidades e não apenas entre EM, impossibilitando assim, a utilização do corpus de avaliação do HAREM e a comparação directa com os sistemas participantes.

O corpus de avaliação de textos para extracção de relações foi constituído a partir de dez manuais escolares da disciplina de Português ao nível do ensino secundário, tendo sido utilizados dois manuais do 10º ano de escolaridade, quatro do 11º e quatro do 12º. Os textos foram seleccionados manualmente através da leitura dos mesmos, de forma a garantir a existência de relações familiares, referências a períodos de vida e à localização de pessoas. Note-se que essa pesquisa não foi feita sobre o conteúdo integral dos manuais, mas apenas sobre determinados tipos de texto. Assim, por se considerar escassa a presença deste tipo de relações em textos poéticos ou dramáticos, por exemplo, privilegiaram-se os textos narrativos e biográficos. O corpus de avaliação é, assim, constituído por 110 textos, com 40,305 palavras.

A distribuição de relações existentes no corpus de avaliação está presente na tabela 5.1. Ao todo, foram identificadas 599 relações. Como se pode verificar alguns tipos de relações não se encontram presentes no corpus (exemplo: *BUILDING-LOCATION* e *CLIENT*).

A tabela 5.1 apresenta o número total de ocorrências de uma determinada categoria, sendo

depois discriminada em percentagem para os diferentes subtipos dessa mesma categoria.

Tabela 5.1: Distribuição das relações

Relação	N.º de Ocorrências
FAMILY	205
LIFETIME	105
BORN	51.43%
DEATH	48.57%
PEOPLE-LOCATION	144
RESIDENCE	16.67%
PLACE-OF-BIRTH	40.28%
COUNTRY-OF-BIRTH	18.06%
PLACE-OF-DEATH	25%
BUILDING-LOCATION	0
BUSINESS	145
EMPLOYEE	11.03%
PROFESSION	82.07%
FOUNDER	2.76%
CLIENT	0%
OWNER	1.38%
GLOBAL-AFFILIATION	2.76%

5.1 Métricas

Apresentam-se agora, as três métricas utilizadas na avaliação. Estas medidas são as mesmas que têm sido empregues em avaliações similares, nomeadamente no HAREM, e são as mais usuais em PLN.

```
Precisão = Relações Correctas / Relações Identificadas

Abrangência = Relações Correctas / Total de Relações

Medida-F = (2 * Precisão * Abrangência) / (Precisão + Abragência)
```

A avaliação processa-se de dois modos: uma avaliação estrita (*aval1*) e outra mais relaxada (*aval2*).

Considera-se que na avaliação estrita (*aval1*) uma relação está correcta quando o tipo de relação e os argumentos da mesma são precisamente iguais à anotação manual feita no corpus de avaliação.

No segundo modo de avaliação (aval2) não se penalizou a extracção de relações pelos erros ocorridos em módulos anteriores, como na identificação de entidades mencionadas e na construção dos nós. O objectivo desta segunda avaliação é quantificar o impacto que estas tarefas têm no desempenho do módulo de extracção de relações.

5.2 Resultados.

Nesta secção apresentam-se os resultados obtidos pelo sistema de extracção de relações. Analisou-se a saída do sistema comparando-a com a anotação manual feita ao mesmo corpus, fazendo duas avaliações simultaneamente.

Na tabela 5.2 apresentam-se os resultados globais obtidos para ambas as avaliações.

Tabela 5.2: Resultados Globais

	Precisão	Abrangência	Medida-F
Avaliação 1	0.36	0.11	0.16
Avaliação 2	0.64	0.22	0.33

Os resultados da primeira avaliação são bastante baixos, o que se compreende, visto que a tarefa de extracção de relações é uma das últimas etapas na cadeia de processamento e todas as falhas ocorridas anteriormente se reflectem aqui.

Concretamente, verificam-se falhas no reconhecimento de entidades mencionadas; na construção de nós, isto é, nos constituintes básicos (*chunks*) das frases; na extracção de dependências entre esses nós, de que resulta em grande medida a interpretação das frases; entre outras, que, no seu conjunto, têm um impacto directo e apreciável na tarefa de extracção de relações.

A segunda avaliação apresenta resultados quase duas vezes superiores à primeira avaliação. São sobretudo as falhas nos módulos de identificação de entidades mencionadas e na construção de nós que causam uma redução de quase 50% no desempenho do módulo de extracção de relações.

Estes resultados não divergem, pois, dos que se verificaram no exercício de avaliação conjunta realizado no HAREM, ainda que não possam ser directamente comparados. Apesar de baixos, os resultados herdam a complexidade da tarefa e os problemas resultantes da construção de raiz de um módulo de extracção de relações.

De seguida, apresentam-se os resultados discriminados para cada uma das categorias apresentadas no capítulo 4. Analisam-se os resultados obtidos bem como os problemas e dificuldades encontrados que impediram a obtenção de melhores resultados.

5.2.1 FAMILY

Na tabela 5.3 estão os resultados obtidos para a relação *FAMILY*. Como neste caso os subtipos são apenas diferentes tipos de relações familiares, decidiu-se não discriminar os resultados parciais.

Tabela 5.3: Avaliação Relação FAMILY

	Precisão	Abrangência	Medida-F
FAMILY	-	-	-
Av. 1	0.32	0.11	0.17
Av. 2	0.49	0.22	0.31

Os resultados obtidos na primeira avaliação das relações familiares são bastante baixos, estando dentro dos valores globais alcançados.

Para a segunda avaliação os resultados melhoraram, apesar de a melhoria não ser tão considerável como na avaliação global. Isso deveu-se ao facto de nas relações familiares existirem dificuldades em reconhecer padrões de frases como por exemplo, frases escritas na forma passiva, frases com ausência de sujeito ou quando o sujeito não é identificado pelos módulos anteriores, entre outros.

De seguida, apresentam-se os principais problemas que influenciaram negativamente o desempenho na extracção de relações familiares.

Um dos problemas encontrados na avaliação da categoria *FAMILY* está relacionado com falhas na detecção de entidades mencionadas, como por exemplo, na identificação de nomes próprios estrangeiros como entidades *"humanas"*.

A grande maioria das relações familiares existentes no corpus de avaliação verifica-se entre entidades da mitologia grega (exemplo: Júpiter, Vénus, Marte, etc.), ora, como estes nomes se encontram identificados como corpos celestes, e não se tinha ainda considerado o seu emprego como entidades mitológicas, tal prejudicou o desempenho do módulo de extracção de relações entre este tipo de entidades.

Na construção de nós, verificam-se casos em que o nome de uma entidade fica em dois nós e é analisado como se fossem duas entidades distintas. Por exemplo a seguinte entidade *"D. Maria Álvares"* foi identificada como duas entidades do tipo *"humana"* (*"D. Maria"* e *"Álvares"*).

A dificuldade em reconhecer se o substantivo que representa o tipo da relação está também a representar uma entidade *"humana"* prejudica duplamente o desempenho do sistema.

Por exemplo, na seguinte frase: *"Dizendo chamar-se Jasão, o filho de Aéson que fora criado por Quíron, reclamava o trono de Iolcos, que era seu por direito."* o sistema não fez a associação da entidade *"Jasão"* à palavra *"filho"*, então considera erradamente a palavra *"filho"* como uma das entidade humanas da relação. Assim, a abrangência diminui, pois a relação correcta não é extraída, e a precisão também diminui porque, é extraída uma relação inexistente.

5.2.2 LIFETIME

A tabela 5.4 apresenta os resultados obtidos para a relação *LIFETIME* e para os subtipos *BORN* e *DEATH*.

Tabela 5.4: Avaliação Relação LIFETIME

	Precisão	Abrangência	Medida-F
LIFETIME	-	-	-
Av. 1	0.63	0.14	0.23
Av. 2	0.95	0.20	0.33
BORN	-	-	-
Av. 1	0.67	0.19	0.29
Av. 2	1	0.24	0.39
DEATH	-	-	-
Av. 1	0.56	0.10	0.17
Av. 2	0.89	0.16	0.27

Os resultados obtidos na primeira avaliação para a categoria *LIFETIME* são consideravelmente baixos. A resolução dos problemas na identificação de entidades mencionadas permitiu um aumento de 0.3 na precisão, de 0.06 na abrangência e de 0.10 na medida-F. Tal demonstra o peso destas tarefas de Reconhecimento de EM (REM) no desempenho do sistema de extracção de relações.

O principal motivo que levou a um valor mais baixo na abrangência consiste na não identificação de expressões temporais absolutas. Por exemplo na seguinte frase: *"O Daniel nasceu em*

1986." *"1986"* foi identificado como *"quantidade"* e não como *"tempo absoluto"*, no entanto, se a frase for: *"O Daniel nasceu no ano de 1986."* *"1986"* já é identificado como expressão temporal absoluta. Ora no corpus de avaliação, abundam biografias onde apenas há informação sobre o ano de nascimento ou o ano de morte sem o respectivo nome classificador (*"ano"*). Como essas expressões temporais não são identificados correctamente, a relação também não é extraída, levando assim a uma diminuição na abrangência do sistema.

Esta situação é o resultado do facto de o módulo de reconhecimento de expressões temporais se encontrar em fase de profunda reestruturação (Maurício, 2011), o que prejudicou o seu desempenho. Recorde-se que na tarefa de REM de tempo, o sistema obteve os melhores resultados de entre todos os participantes do HAREM. Espera-se, pois, que no fim do processo de reformulação, desse módulo, o sistema produza já resultados consideravelmente melhores na extracção de relações que envolvem expressões temporais.

5.2.3 PEOPLE-LOCATION

Na tabela 5.5 encontram-se os resultados obtidos para a relação *PEOPLE-LOCATION* e para os subtipos *RESIDENCE*, *PLACE-OF-BIRTH*, *COUNTRY-OF-BIRTH* e *PLACE-OF-DEATH*.

Tabela 5.5: Avaliação Relação PEOPLE-LOCATION

	Precisão	Abrangência	Medida-F
PEOPLE-LOCATION	-	-	-
Av. 1	0.71	0.15	0.25
Av. 2	0.92	0.24	0.38
RESIDENCE	-	-	-
Av. 1	0.5	0.04	0.07
Av. 2	0.67	0.08	0.15
PLACE-OF-BIRTH	-	-	-
Av. 1	0.75	0.26	0.38
Av. 2	0.95	0.34	0.51
COUNTRY-OF-BIRTH	-	-	-
Av. 1	1	0.08	0.14
Av. 2	1	0.12	0.21
PLACE-OF-DEATH	-	-	-
Av. 1	0.57	0.11	0.19
Av. 2	0.90	0.25	0.39

Tal como nas categorias anteriores, *PEOPLE-LOCATION* apresenta melhorias significativas

na segunda avaliação, onde as subidas na medida-F variam entre 0.07 e 0.2. Essa subida deve-se sobretudo à correcta identificação das entidades do tipo *"humano"*.

Um dos principais problemas encontrados consiste nos casos com expressões que requerem a resolução de correferência, como por exemplo, na frase: "O João nasceu em Lisboa e morreu na mesma cidade". Aqui, é necessário determinar que a expressão *"na mesma cidade"* se refere à entidade *"Lisboa"*. Como essa correferência não foi estabelecida, não foram extraídas as correspondentes relações.

Verifica-se ainda outra situação, quando um dado lugar é referido não apenas por uma expressão mas por uma sucessão de expressões progressivamente mais genéricas (ou mais específicas), como por exemplo: *"Fernando António Nogueira Pessoa nasceu em Lisboa, no Largo de São Carlos no dia 13 de Junho de 1888."* onde inicialmente se refere uma localização genérica *"Lisboa"* e depois uma mais específica *"Largo de São Carlos"*. Nestes casos, o sistema apenas extrai a relação com a primeira localização. A solução para estes casos passaria por validar uma ontologia geográfica para as relações de hiponímia (inclusão) dos termos em aposição e extrair as relações em falta. Tal, porém, ultrapassa o âmbito desta dissertação.

5.2.4 BUILDING-LOCATION

Como no corpus de avaliação não existem casos que representem uma relação de *"Localização de Empresas"*, não se pode retirar conclusões sobre o trabalho desenvolvido nesta categoria. Pretende-se avaliar o sistema neste aspecto quando se utilizar um novo corpus de avaliação entretanto já compilado e em fase de anotação.

5.2.5 BUSINESS

Na tabela 5.6 apresentam-se os resultados obtidos para a relação *BUSINESS* e para os respectivos subtipos *EMPLOYEE*, *PROFESSION*, *FOUNDER*, *CLIENT*, *OWNER* e *AFFILIATION*.

A diferença entre os resultados obtidos em cada avaliação é muito grande, com valores dez vezes superiores em todas as métricas. O principal factor para esta discrepância de resultados é uma falha na construção dos nós, o que já não foi contabilizada na segunda avaliação.

Tabela 5.6: Avaliação Relação BUSINESS

	Precisão	Abrangência	Medida-F
BUSINESS	-	-	-
Av. 1	0.06	0.02	0.03
Av. 2	0.62	0.20	0.30
EMPLOYEE	-	-	-
Av. 1	0	0	0
Av. 2	0	0	0
PROFESSION	-	-	-
Av. 1	0.07	0.03	0.04
Av. 2	0.61	0.24	0.34
FOUNDER	-	-	-
Av. 1	0	0	0
Av. 2	0	0	0
OWNER	-	-	-
Av. 1	0	0	0
Av. 2	0	0	0
AFFILIATION	-	-	-
Av. 1	0	0	0
Av. 2	1	0.25	0.40

Como exemplo, na seguinte frase: *"O rei D. Dinis era conhecido como o Lavrador."* a relação extraída é *BUSINESS_CARGO_PROFESSION(rei D. Dinis,rei)*, mas a relação correcta é *BUSINESS_CARGO_PROFESSION(D. Dinis,rei)*. Na segunda avaliação não se contabilizaram estes casos como errados, o que permitiu uma melhoria considerável no desempenho.

Como essa construção de nós extraía uma relação errada, o sistema era penalizado na precisão, na abrangência e consequentemente na medida-F. A correcta identificação das entidades *"humanas"* também permitiu aumentar o desempenho desta relação.

Nas restantes subcategorias os resultados são pouco significativos, visto que o número de ocorrências deste tipo de relação no corpus de avaliação é bastante reduzido. Dessa forma não se retiram conclusões sobre as relações *CLIENT*, *FOUNDER*, *OWNER* e *AFFILIATION*.

Quanto à relação *EMPLOYEE* os resultados foram decepcionantes e derivam da não identificação de entidades que representam uma empresa ou uma organização. Quando não é feita essa identificação então também não é feita a extracção da relação correspondente. Ora, a identificação de empresas e organizações, no HAREM, foi justamente um dos tipos de EM em que o sistema apresenta resultados não tão satisfatórios como os que obteve noutras categorias, como por exemplo, nas expressões temporais. Espera-se, pois, que a melhoria do desempenho

do módulo REM nestas categorias (Oliveira, 2010) possa repercutir-se no módulo de extracção de relações.

Verificou-se ainda uma situação concreta que prejudicou o desempenho do sistema: sempre que o nome de uma pessoa aparecia associado a um conjunto de profissões uma enumeração, o sistema extraía erradamente relações de profissão entre os diferentes itens da enumeração. Assim por exemplo, na frase: *"O João é cantor, músico e actor."* a entidade *"João"* está associada a três nomes de profissão. Na gramática do XIP, todos os nomes de profissão têm o traço *"human"* pois aplicam-se e representam entidades *"humanas"*. Ora, neste caso, o sistema assume que *"músico"* está a representar uma entidade e não o atributo *"profissão"* de outra entidade, no caso o *"João"*. Por essa razão são incorrectamente extraídas relações de profissão entre os outros nomes da enumeração, como por exemplo *BUSINESS_PROFESSION(músico, actor)*. Este problema é resultado das dificuldades no tratamento da coordenação, que neste momento é apenas resolvida a nível estritamente local (coordenação de dois termos). A propagação da dependência atributo entre o sujeito e o primeiro item da enumeração aos restantes itens coordenados deveria permitir resolver correctamente a extracção de relações nesta situação sintáctica.

5.3 Nova Avaliação

Com a conclusão do trabalho desenvolvido por (Oliveira, 2010), na identificação e reconhecimento de entidades mencionadas, decidiu-se proceder a uma nova avaliação para quantificar as melhorias na tarefa de extracção de relações. Os novos resultados encontram-se na tabela 5.7.

Tabela 5.7: Novos Resultados Globais

	Precisão	Abrangência	Medida-F
Avaliação 1	0.46	0.11	0.18
Avaliação 2	0.70	0.22	0.34

Os novos resultados foram ligeiramente melhores sobretudo na precisão que subiu 0.1 na *Avaliação 1* e 0.6 na *Avaliação 2*, mas, no entanto, os resultados da primeira avaliação continuam a ser penalizados por não reconhecer os nomes mitológicos como entidades humanas. Decidiu-se então, fazer uma nova avaliação com um corpus diferente do primeiro, de forma a abranger não só outro tipo de relações como também diferentes tipos de texto.

Este corpus de textos para extracção das relações foi compilado a partir de textos disponíveis em linha, utilizando a pesquisa avançada do Google e restringindo a procura a textos de páginas localizadas em Portugal, escritos em português e sob o domínio de topo '.pt'. Para os padrões de busca utilizaram-se sequências de palavras do mesmo campo semântico. Assim, para as relações profissionais seleccionaram-se textos contendo todas as palavras da série, por exemplo: profissional, profissão; trabalhador, empregado, chefe, director, gerente, gestor; empresa, empresário; accionista, proprietário, vender, comprar, adquirir; cliente, clientela, entre outros.

A selecção dos textos baseou-se na leitura cursiva dos textos, de forma a verificar se alguma das relações-alvo estava presente e, se possível, se a densidade dessas relações os tornava interessantes para esta tarefa. Os textos foram copiados *in totum* e a ligação registada, com a respectiva data de acesso. O segundo corpus contém 33 textos, com 16.156 palavras, recolhidos entre 11 e 19 de Setembro de 2009.

Na tabela 5.8 apresentam-se os resultados globais obtidos para este segundo corpus.

Tabela 5.8: Resultados Globais 2º Corpus

	Precisão	Abrangência	Medida-F
2º Corpus	0.63	0.22	0.32

Durante esta fase de avaliação introduziram-se algumas alterações na cadeia de processamento, provocando assim modificações nas árvores sintácticas construídas pelo XIP. Essas mesmas alterações também se reflectem nos argumentos das relações. Por exemplo, na frase *"Esmeralda Dourado - Presidente da Comissão Executiva da SAG Gest."*, a expressão *"Presidente da Comissão Executiva"* era um único nó mas, neste momento, está dividido em dois nós, *"Presidente"* e *"da Comissão Executiva"*. Assim, a anotação manual tem a relação *BUSINESS_PROFESSION(Esmeralda Dourado, Presidente da Comissão Executiva)* enquanto o sistema extrai a relação *BUSINESS_PROFESSION(Esmeralda Dourado, Presidente)*. Optou-se por não contabilizar estes casos como incorrectos, visto que, apesar de o nome do argumento ser diferente, ambas as anotações se referem à mesma entidade.

Os resultados globais, obtidos para esta avaliação, estão próximos da *Avaliação 2* feita ao 1º corpus mas, neste caso, as falhas no REM não foram minimizadas, prejudicando o desempenho da tarefa de extracção de relações.

6 Conclusão e Trabalho Futuro

A extracção de relações entre entidades não é uma tarefa simples. A complexidade das línguas naturais dificulta a extracção automática de toda a informação presente nos textos e a consequente compreensão do seu conteúdo, que poderia ser útil para inúmeras aplicações na sociedade contemporânea.

Neste trabalho, decidiu-se constituir um conjunto de directivas de anotação e utilizar um corpus de avaliação especificamente concebido para este efeito. Tal decisão tem por base a opção de considerar, para efeitos de extracção de relações, não apenas as que envolvem entidades mencionadas mas também outras entidades, normalmente não abrangidas pela tarefa de REM e que são expressas através de léxico comum. Ora, esta opção diverge substancialmente da que foi tomada pela organização do Segundo HAREM, o evento de avaliação conjunta dos sistemas de processamento de língua portuguesa em que primeiro se tentou uma avaliação comparativa de sistemas de extracção de relações. Por outro lado, o conjunto de categorias consideradas no âmbito da tarefa de REM também difere em alguns aspectos não triviais (Oliveira, 2010, Maurício, 2011).

Por essa razão, não foi possível utilizar o corpus de avaliação constituído para aquela campanha de avaliação conjunta, sendo também difícil a comparação directa do sistema aqui desenvolvido com os sistemas participantes no Segundo HAREM.

Ainda assim, os resultados globais aqui alcançados não se afastam de modo significativo dos resultados desse exercício, no entanto, em ambos os casos, os resultados estão longe de poderem ser considerados plenamente satisfatórios. Em particular, a principal dificuldade resulta do facto de a extracção de relações ter lugar nas etapas finais do processamento de texto, pelo que sofre com a sucessiva acumulação de erros ou omissões resultantes dos módulos anteriores da cadeia de processamento.

Por essa razão, decidiu-se proceder a uma avaliação dos resultados em dois modos distintos: Uma primeira avaliação mais estrita, e uma segunda avaliação em que algumas falhas dos

módulos anteriores não são contabilizadas.

Dos resultados pode depreender-se que o desempenho do sistema se assemelha à de outros sistemas idênticos já avaliados para o português, revelando a complexidade e a dificuldade da tarefa de extracção de relações.

Finalmente, os resultados poderiam ter sido mais satisfatórios se já estivessem disponíveis outros trabalhos que estão neste momento a decorrer e que pretendem melhorar o desempenho da cadeia de processamento do L2F em algumas tarefas de que depende a extracção de relações. Trata-se concretamente do reconhecimento e classificação de entidades mencionadas (Oliveira, 2010), resolução de correferência e, em particular, de expressões anafóricas (Nobre, 2010) e ainda da identificação, classificação e normalização de expressões temporais (Maurício, 2011).

É portanto, expectável que o desempenho global do sistema e em particular do módulo de extracção de relações venha a melhorar no futuro, com a integração destes trabalhos.

Existem várias tarefas que poderão melhorar o desempenho actual do sistema de extracção de relações. Ficam aqui algumas sugestões de trabalho futuro.

- Utilizar um corpus maior e mais abrangente, de forma a cobrir todas as categorias de relações e em que cada uma apresente um número considerável de ocorrências. Essa avaliação deverá permitir tirar conclusões mais seguras sobre a qualidade do desempenho do sistema.

- Desenvolver um modelo de avaliação que permita apreciar exclusivamente a tarefa de extracção de relações, evitando assim que esta seja penalizada pelas falhas ocorridas nos módulos anteriores da cadeia de processamento e que são independentes.

- Aumentar a abrangência do sistema, que aqui apresentou valores relativamente baixos, nomeadamente considerando novos padrões, mais complexos dos que actualmente estão descritos.

- Restringir mais as regras já implementadas, tentando minimizar o número de relações extraídas incorrectamente (sobregeração), nomeadamente nos casos das relações familiares e na relação de profissão.

- Alargar o âmbito do conjunto de relações tratadas pelo sistema, tais como relações de autoria e personagens de obras (literárias ou cinematográficas).

- Complementar as relações familiares acrescentando a dimensão temporal/aspectual, nomeadamente indicando se uma determinada relação está em curso ou já terminou, principalmente para a relação de *SPOUSE*.

- Desenvolver um léxico geográfico, ontologicamente estruturado, com todas as regiões portuguesas, permitindo assim deduzir a nacionalidade de uma pessoa caso esta tenha nascido em qualquer parte do território português. É eventualmente desejável utilizar ontologias geográficas já existentes.

Referências

[Agichtein and Gravano, 2000] Agichtein, E. and Gravano, L. (2000). Snowball: Extracting Relations from Large Plain-Text Collections. In *In Proceedings of the 5th ACM International Conference on Digital Libraries (ACM DL)*, pages 85–94.

[Baptista et al., 2006] Baptista, J., Batista, F., and Mamede, N. (2006). Building a Dictionary of Anthroponyms. In *In PROPOR 2006 - Computational Processing of the Portuguese Language, vol. 3960*, pages 21–30. Springer Verlag.

[Bick, 2000] Bick, E. (2000). *The Parsing System Palavras. Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Universidade de Arhus.

[Bruckschen et al., 2008] Bruckschen, M., Camargo de Souza, J., Vieira, R., and Rigo, S. (2008). Sistema SERELEP para o reconhecimento de relações entre entidades mencionadas. In *Desafios na avaliacao conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, chapter 14, pages 247–260.

[Cardoso, 2008] Cardoso, N. (2008). REMBRANDT - Reconhecimento de Entidades Mencionadas Baseado em Relações e ANálise Detalhada do Texto. In *Desafios na avaliacao conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, chapter 11, pages 195–211.

[Cardoso and Santos, 2007] Cardoso, N. and Santos, D. (2007). *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área.*

[Chinchor, 1997] Chinchor, N. (1997). MUC-7 Named Entity Task Definition Dry Run Version. In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*. Morgan Kaufmann Publishers, Inc.

[Culotta and Sorensen, 2004] Culotta, A. and Sorensen, J. (2004). Dependency tree kernels for relation extraction. In *ACL '04: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 423, Morristown, NJ, EUA. Association for Compu-

tational Linguistics.

[Diniz, 2010] Diniz, C. (2010). *RuDriCo 2 - Um Conversor Baseado em Regras de Transformação Declarativas*. Master's thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa, Lisboa, Portugal.

[Doddington et al., 2004] Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., and Weischedel, R. (2004). The Automatic Content Extraction (ACE) Program–Tasks, Data, and Evaluation. *Proceedings of LREC 2004*, pages 837–840.

[Freitas et al., 2009] Freitas, C., Santos, D., Mota, C., Oliveira, H. G., and Carvalho, P. (2009). Relation detection between named entities: report of a shared task. In *DEW '09: Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, pages 129–137, Morristown, NJ, EUA. Association for Computational Linguistics.

[Grishman and Sundheim, 1996] Grishman, R. and Sundheim, B. (1996). Message Understanding Conference-6: a brief history. In *Proceedings of the 16th conference on Computational linguistics*, pages 466–471, Morristown, NJ, USA. Association for Computational Linguistics.

[Jurafsky and Martin, 2000] Jurafsky, D. and Martin, J. H. (2000). *Speech and Language Processing*. Prentice-Hall.

[Loureiro, 2007] Loureiro, J. (2007). *Reconhecimento de Entidades Mencionadas e Normalização de Expressões Temporais*. Master's thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa, Lisboa, Portugal.

[Mamede, 2007] Mamede, N. (2007). *A Cadeia de Processamento XIP*. L^2F – Laboratório de Sistemas de Língua Falada, INESC-ID, Lisboa.

[Maurício, 2011] Maurício, A. (2011). *Identificação, Classificação e Normalização de Expressões Temporais*. Master's thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa, Portugal, Lisboa, Portugal.

[Medeiros, 1995] Medeiros, J. C. (1995). *Processamento Morfológico e Correcção Ortográfica do Português*. Master's thesis, Instituto Superior Técnico – Universidade Técnica de Lisboa, Lisboa, Portugal.

[Miller et al., 1998] Miller, S., Crystal, M., Fox, H., Ramshaw, L., Schwartz, R., Stone, R., Weischedel, R., and Group, T. A. (1998). Algorithms that Learn to Extract Information

BBN: Description of the Sift System as used for MUC-7. In *In Proceedings of MUC-7*.

[Mota and Santos, 2008] Mota, C. and Santos, D. (2008). *Desafios na avaliacao conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*.

[Ng and Cardie, 2002] Ng, V. and Cardie, C. (2002). Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution. In *Proceedings of the 19th international conference on Computational linguistics*, Morristown, NJ, USA. Association for Computational Linguistics.

[Nobre, 2010] Nobre, N. (2010). *Resolução de expressões anafóricas*. Master's thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa, Lisboa, Portugal.

[Oliveira, 2010] Oliveira, D. (2010). *Extraction and Classification of Named Entities*. Master's thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa, Lisboa, Portugal.

[Orasan et al., 2008] Orasan, C., Cristea, D., Mitkov, R., and Branco, A. (2008). Anaphora Resolution Exercise: An overview. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, Paris, France.

[Pardal, 2007] Pardal, J. P. (2007). *Manual do Utilizador do RuDriCo*. L^2F – Laboratório de Sistemas de Língua Falada, INESC-ID, Lisboa.

[Ribeiro et al., 2003] Ribeiro, R., Mamede, N., and Trancoso, I. (2003). Using Morphossyntactic Information in TTS Systems: comparing strategies for European Portuguese. In *Computational Processing of the Portuguese Language: 6th International Workshop, PROPOR 2003, Faro, Portugal, June 26-27, 2003. Proceedings*, volume 2721 of *Lecture Notes in Computer Science*. Springer.

[Romão, 2007] Romão, L. (2007). *Reconhecimento de Entidades Mencionadas em Língua Portuguesa*. Master's thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa, Lisboa, Portugal.

[Roth and Yih, 2002] Roth, D. and Yih, W. (2002). Probabilistic reasoning for entity & relation recognition. In *Proceedings of the 19th international conference on Computational linguistics*, Morristown, NJ, USA. Association for Computational Linguistics.

[Roth and Yih, 2007] Roth, D. and Yih, W. (2007). Global Inference for Entity and Relation Identification via a Linear Programming Formulation. In *Introduction to Statistical Relational Learning*. MIT Press.

[Silveira Chaves, 2008] Silveira Chaves, M. (2008). Geo-ontologias e padrões para reconhecimento de locais e de suas relações em textos: o SEI-Geo no Segundo HAREM. In *Desafios na avaliacao conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, chapter 13, pages 231–245.

[Silveira Chaves et al., 2005] Silveira Chaves, M., J. Silva, M., and Martins, B. (2005). *GKB - Geographic Knowledge Base*. Technical report, Faculdade de Ciências da Universidade de Lisboa.

[Xerox, 2003a] Xerox, R. C. E. (2003a). *Xerox Incremental Parser – Reference Guide*.

[Xerox, 2003b] Xerox, R. C. E. (2003b). *Xerox Incremental Parser – User's Guide (Scripting)*.

[Xerox, 2003c] Xerox, R. C. E. (2003c). *XIP User Guide*.

[Zelenko et al., 2003] Zelenko, D., Aone, C., and Richardella, A. (2003). Kernel Methods for Relation Extraction. In *Journal of Machine Learning Research*, volume 3, pages 1083–1106.

[Zhao and Grishman, 2005] Zhao, S. and Grishman, R. (2005). Extracting relations with integrated information using kernel methods. In *ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 419–426, Morristown, NJ, EUA. Association for Computational Linguistics.

A Directivas

Este documento descreve as directivas adoptadas para a classificação e anotação de relações extraídas a partir de textos.

As relações são representadas como dependências entre um conjunto de argumentos.

```
<RELACAO>_<TIPO>_<features> (<arg1>,<arg2>)
```

Os traços incluem, por exemplo, o género (M/F) dos argumentos (1/2), quando tal informação é pertinente e/ou está disponível na expressão analisada.

De um modo geral, as relações estabelecem-se não apenas entre Entidades Mencionadas (EM) (e.g. *"O Pedro vive em Lisboa"*) mas também entre outras expressões designativas dos diferentes tipos de entidades tratadas pelo sistema (*"O João é lisboeta"*).

```
LOCATION_RESIDENCE (Pedro, Lisboa)
LOCATION_RESIDENCE (Pedro, lisboeta)
```

Além disso, no caso de relações explícitas em que um dos seus argumentos se encontra omisso (e.g. *"O Pedro nasceu em Lisboa mas vive actualmente no Porto"*), ou é indirectamente referido pelo próprio nome da relação (*"O pai do Pedro disse-lhe isso"*), adopta-se a estratégia de representar a relação, ainda que incompleta, indicando no lugar do argumento elidido/subentendido um pronome pessoal, cujo género e número é indicado na medida em que essa informação estiver disponível.

Assim, para os exemplos acima, extraem-se as relações:

```
LOCATION_PLACE-OF-BIRTH (Pedro, Lisboa)
LOCATION_RESIDENCE (Pedro, lisboeta)
FAMILY_PARENT_1M_2M (Ele, Pedro)
```

No caso das relações assimétricas (ou orientadas), a ordem dos argumentos é relevante. É o caso da relação de parentesco 'pai', acima, pelo que na representação o ascendente figura como primeiro argumento. No caso de relações simétricas, p.ex. 'irmão', a ordem dos argumentos é aquela pela qual estes aparecem expressos linearmente no texto.

A.1 Relações familiares

As relações familiares capturam o parentesco entre duas pessoas e são representadas pela dependência *FAMILY* (família). Cada grau de parentesco corresponde a um tipo distinto. Consideram-se por ora os seguintes tipos:

A.1.1 Relações Assimétricas

- *PARENT* (entre progenitor (pai) e descendente (filho); abrange igualmente relações entre padrasto/enteado e entre pai/filho adoptivo);
- *GRANDPARENT* (entre progenitor (avô) e descendente (neto), inclui também relações de (bisavô) e (trisavô));
- *UNCLE* (tio/sobrinho);
- *FATHER-IN-LAW* (sogro/genro e nora);
- *GODFATHER* (padrinho/afilhado);
- *GRAND-UNCLE* (tio-avô/sobrinho-neto, inclui também relações de tio-bisavô e tio-trisavô);

A.1.2 Relações Simétricas

- *SIBLING* (irmãos),
- *SPOUSE* (pessoas casadas ou em união de facto)
- *COUSIN* (primos)
- *BROTHER-IN-LAW* (cunhados)

De seguida, apresenta-se um exemplo para cada subtipo aqui referido, bem como a notação usada pelo sistema

- *O João é pai do Pedro*

  `FAMILY_PARENT_1M_2M(João, Pedro)`

- *O João é neto do Pedro*

  `FAMILY_GRANDPARENT_1M_2M(Pedro, João)`

- *O João é tio do Pedro*

  `FAMILY_UNCLE_1M_2M(João, Pedro)`

- *O João é sogro do Pedro*

  `FAMILY_FATHER-IN-LAW_1M_2M(João, Pedro)`

- *O João é padrinho do Pedro*

  `FAMILY_GODFATHER_1M_2M(João, Pedro)`

- *O João é tio-avô do Pedro*

  `FAMILY_GRAND-UNCLE_1M_2M(João, Pedro)`

- *A Joana é irmã do Pedro*

  `FAMILY_SIBLING_1F_2M(Joana, Pedro)`

- *A Joana é esposa do Pedro*

  `FAMILY_SPOUSE_1F_2M(Joana, Pedro)`

- *A Joana é prima do Pedro*

  `FAMILY_COUSIN_1F_2M(Joana, Pedro)`

- *A Joana é cunhada do Pedro*

  `FAMILY_BROTHER-IN-LAW_1F_2M(Joana, Pedro)`

A.2 Período de Vida

Estas relações são representadas pela dependência *LIFETIME*, que associa uma pessoa à sua data de nascimento e de óbito, representadas, respectivamente, pelos subtipos *BORN* e *DEATH*. As datas são expressões temporais (TIMEX) absolutas (REF) e são representadas pela expressão do texto, sem normalização, podendo apresentar dia-mês-ano, mês-ano e ano.

Exemplos:

- *O João nasceu em 21 de Janeiro de 1918*

  `LIFETIME_BORN(João, 21 de Janeiro de 1918)`

- *O João morreu em Dezembro de 1990*

  `LIFETIME_DEATH(João, Dezembro de 1990)`

A.3 Localização de Pessoas

Estas relações são representadas pela dependência *PEOPLE-LOCATION*, que associa uma pessoa a uma localização. Inclui os subtipos *RESIDENCE* (residência/morada), *PLACE-OF-BIRTH* (local de nascimento), *COUNTRY-OF-BIRTH* (país de nascimento) e *PLACE-OF-DEATH* (local de morte). Como os nomes indicam, para a naturalidade distinguem-se os nomes de países, que podem frequentemente ser associados à nacionalidade, dos outros nomes de locais.

Exemplos

- *O João reside na Avenida da Liberdade*

  `PEOPLE-LOCATION_RESIDENCE(João, Avenida da Liberdade)`

- *O João nasceu em Lisboa*

  `PEOPLE-LOCATION_PLACE-OF-BIRTH(João, Lisboa)`

- *O João é português*

  `PEOPLE-LOCATION_COUNTRY-OF-BIRTH(João, português)`

- *O João morreu no Porto*

  `PEOPLE-LOCATION_PLACE-OF-DEATH(João, Porto)`

A.4 Localização de Edifício

Esta relação, representada pela dependência *BUILDING-LOCATION*, semelhante às anteriores, associa, instituições ou monumentos a locais:

Exemplos:

- *A Torre de Belém situa-se em Lisboa*

  `BUILDING-LOCATION(Torre de Belém, Lisboa)`

- *A sede da Microsoft é nos Estados Unidos.*

  `BUILDING-LOCATION(Microsoft, Estados Unidos)`

A.5 Relações Empresariais

As relações empresariais são captura das pela dependência *BUSINESS* e relacionam uma pessoa com uma instituição, empresa, profissão, religião, etc. Inclui os seguintes tipos:

- *EMPLOYEE*: relação entre pessoa e empresa, independentemente da função que desempenha;
- *PROFESSION*: profissão ou função de uma pessoa; um traço *CARGO* é adicionado nos casos em que há identidade entre cargo e função/profissão;
- *FOUNDER*: pessoa fundadora de empresa;
- *CLIENT*: relação entre pessoa e empresa de que aquela é cliente;
- *OWNER*: proprietário de empresa, independentemente da natureza da propriedade (dono, sócio, accionista, etc.)
- *AFFILIATION*: relação entre pessoa e instituição ou outro tipo de entidade (não uma empresa) na que aquela se encontra afiliada; estas entidades podem ser de natureza desportiva, religiosa, associativa, recreativa, nacionais e internacionais, etc.

Apresenta-se agora, um exemplo para cada tipo de relação descrito nesta secção.

- *O João trabalha na Microsoft*

  `BUSINESS_EMPLOYEE(João, Microsoft)`

- *O João é arquitecto*

  `BUSINESS_PROFESSION(João, arquitecto)`

- *Bill Gates fundou a Microsoft*

  `BUSINESS_FOUNDER(Bill Gates, Microsoft)`

- *O João é cliente da HP*

  `BUSINESS_CLIENT(João, HP)`

- *O João é accionista da EDP*

  `BUSINESS_OWNER(João, EDP)`

- *O João é socialista*

  `BUSINESS_AFFILIATION(João, socialista)`