

---

# Sistema de Pergunta-Resposta em área médica

---

DISSERTAÇÃO

apresentada para  
obtenção do grau de

MESTRE

em

ENGENHARIA INFORMÁTICA

pelo autor

Pedro Miguel Rocha Pereira Fialho

tendo como orientador

Paulo Miguel Torres Duarte Quaresma



Departamento de Informática  
Escola de Ciências e Tecnologia,  
Universidade de Évora  
Évora, Portugal  
[www.di.uevora.pt](http://www.di.uevora.pt)



---

# Sistema de Pergunta-Resposta em área médica

---

Autor: Pedro Miguel Rocha Pereira Fialho  
Orientador: Paulo Miguel Torres Duarte Quaresma

## Resumo

A correta implementação de sistemas de pergunta-resposta representa uma mudança no conceito de interação entre Homem e máquina devido à utilização exclusiva de linguagem humana na comunicação em ambos os sentidos. Existem ainda poucas alternativas de sistemas de pergunta-resposta para língua portuguesa, orientados a domínios médicos ou baseados em ontologias e/ou programação em lógica. Esta dissertação descreve a implementação e avaliação de um sistema de pergunta-resposta em lógica contextual para interrogação em português de uma ontologia de relatórios neurovasculares definindo uma combinação inovadora de tecnologias em sistemas de pergunta-resposta. Esta abordagem é avaliada com um conjunto de perguntas formulado especificamente para este domínio médico relativamente a precisão, cobertura, acerto e pontuação F revelando resultados satisfatórios/medianos.

**Palavras-Chave** — Sistemas de pergunta-resposta, recuperação de informação, ontologias, programação em lógica contextual.



---

# Question Answering system for a medical domain

## Abstract

The mature development of question answering systems sets a turning point in human-computer interaction due to the exclusive use of human language on both ways of communication. There are still few alternatives for question answering systems for the portuguese language, medical domain oriented or based on ontologies and/or logic programming. This thesis describes the development and evaluation of a question answering system in contextual logic allowing portuguese queries on an ontology of neurovascular reports defining a new combination of technologies in question answering systems. This approach is evaluated with a domain specific question set for precision, recall, accuracy and F-score showing reasonable results.

*Keywords* — Question answering, information retrieval, ontologies, contextual logic programming.



---

# Agradecimentos

Agradeço à minha mãe Maria D'Aires e ao meu pai Joaquim por me acompanharem nesta e noutras etapas da vida, ao meu tio Frederico pelo incentivo e aconselhamento nesta e noutras etapas da vida, ao meu orientador de tese Paulo Quaresma por tornar possível este trabalho, às minhas orientadoras de investigação Irene Rodrigues e Luísa Coheur pela disponibilidade, incentivo e apoio ao desenvolvimento deste e de outros trabalhos, ao professor Salvador Abreu pela disponibilidade e ajuda na obtenção e manuseamento dos sistemas XPTO e GProlog/CX e ao João Laranjinho pela disponibilidade e ajuda na obtenção e manuseamento do analisador sintático/semântico.

Pedro Miguel Rocha Pereira Fialho  
Viana do Alentejo, Portugal  
5 de Fevereiro de 2012



---

# Conteúdo

<b>Agradecimentos</b>	<b>v</b>
<b>Conteúdo</b>	<b>vii</b>
<b>Lista de Figuras</b>	<b>ix</b>
<b>1 Introdução</b>	<b>1</b>
1.1 Motivação . . . . .	1
1.2 Objetivo . . . . .	2
1.3 Conceitos envolvidos . . . . .	2
1.4 Estrutura da dissertação . . . . .	4
<b>2 Sistemas de Pergunta e Resposta</b>	<b>5</b>
2.1 Componentes . . . . .	6
2.1.1 Processamento de Língua Natural . . . . .	6
2.1.2 Representação de conhecimento . . . . .	7
2.2 Arquiteturas . . . . .	8
2.3 Sumário . . . . .	9
<b>3 CliniQA: um SPR na medicina</b>	<b>11</b>
3.1 Componentes base . . . . .	11
3.1.1 Ontologia de Relatórios Neurovasculares . . . . .	12
3.1.2 Analisador sintático e semântico com marcação de nomes próprios .	13
3.1.3 XPTO Prolog Translation of Ontologies . . . . .	15
3.2 Adaptações . . . . .	16
3.3 Arquitetura . . . . .	18
3.3.1 Processamento da pergunta . . . . .	18
3.3.2 Recuperação de informação . . . . .	20
3.3.3 Composição da resposta . . . . .	21
3.4 Sumário . . . . .	26

<b>4</b>	<b>Avaliação</b>	<b>27</b>
4.1	Conjunto de teste . . . . .	27
4.2	Métricas experimentais . . . . .	29
4.2.1	Seleção de factos . . . . .	30
4.2.2	Respostas finais . . . . .	31
4.3	Resultados . . . . .	32
4.3.1	Discussão . . . . .	33
4.4	Sumário . . . . .	35
<b>5</b>	<b>Conclusão</b>	<b>37</b>
5.1	Contributo . . . . .	37
5.2	Limitações . . . . .	38
5.3	Trabalho futuro . . . . .	38
5.4	Sumário . . . . .	40
	<b>Bibliografia</b>	<b>41</b>
<b>A</b>	<b>Perguntas de teste</b>	<b>49</b>

---

## Lista de Figuras

3.1	Representação no Protégé do indivíduo “Diagnóstico28” que pertence simultaneamente às classes “Ligeira”, “Moderada”, “Grave” e “Pré-Oclusiva”, está relacionado com “Observação” e “Artéria” e é descrito por informação do UMLS.	13
3.2	Conjunto com a única DRS da pergunta « <i>Quais os nomes dos pacientes diagnosticados?</i> ». Cada linha representa um constituinte da DRS e argumento do predicado <i>drs/3</i> (por ordem de ocorrência): 1- variáveis/referentes, 2 - factos/condições e 3 - ação/foco. . . . .	14
3.3	Múltiplas DRS da pergunta « <i>Que aparelhos foram utilizados no exame da paciente Ana Maria?</i> » ilustrando o significado inferível do verbo/ação principal por substituição das variáveis pelos respetivos funtores e conjugação do verbo.	14
3.4	Excertos de <i>units</i> geradas com indicação dos respetivos ficheiros Prolog. . . . .	15
3.5	Exemplo de alterações para conformidade com o XPTO. . . . .	17
3.6	Arquitetura global do CliniQA. . . . .	19
3.7	Informação recolhida automaticamente da pergunta “Qual a informação clínica da paciente Paula Rosado?”. . . . .	20
3.8	Contextos recolhidos e respetiva proveniência das palavras inferidas (conjuntos vazios omitidos). . . . .	22
3.9	Excerto do grafo sobre a paciente “Ana Maria”. . . . .	23
3.10	Informação direta sobre a paciente “Ana Maria”. . . . .	23
3.11	Resposta à pergunta “Qual a conclusão do relatório da paciente Vanessa Capucho?” (impressa de uma lista Prolog). O relatório correspondente contém a conclusão “Hipoplasia da Artéria Vertebral Esquerda/Direita”. . . . .	25
4.1	Definição de medida F, comum entre avaliações. . . . .	29
4.2	Definição de cobertura, precisão e acerto para micro avaliação. . . . .	31
4.3	Definição de cobertura, precisão e acerto para macro avaliação. . . . .	31
4.4	Definição de acerto e cobertura excluindo a influência do analisador linguístico.	32
4.5	Valores de precisão, acerto, cobertura e medida F para macro e micro avaliação.	33
4.6	Resposta à pergunta “Qual a conclusão do relatório da paciente Paula Rosado?”. . . . .	34



# Capítulo 1

---

## Introdução

*Motivação, conceitos e objetivo da implementação de Sistemas de Pergunta-Resposta na medicina.*

Encontrar informação é uma tarefa essencial e determinante para a usabilidade de repositórios/representações de conhecimento, sendo a sua complexidade proporcional à quantidade e variedade da informação representada. Os Sistemas de Informação (SI) englobam interfaces (tipicamente) textuais para especificação de palavras (chave) úteis na Recuperação de Informação (RI) pretendida. A forma mais elementar de pesquisa consiste em coincidir estas palavras (por padrão sintático ou relação semântica) com elementos do conhecimento pelo que, a utilidade dos resultados apresentados depende da sua seleção e organização[8].

Formular um conjunto de palavras-chave é uma adaptação humana ao funcionamento da máquina/SI visto que estas provém (essencialmente) da pergunta que surge na língua natural do utilizador (motivação para a pesquisa) perante uma situação de dúvida ou falta de informação. Recorrendo a técnicas para Processamento de Língua Natural (PLN) esta pergunta pode ser interpretada automaticamente permitindo uma especificação explícita e natural da informação pretendida. O PLN pode também ser utilizado para gerar língua natural com o conhecimento selecionado apesar desta tarefa não ser determinante para obtenção de uma resposta. Os mecanismos de pesquisa que permitem filtrar uma representação de conhecimento através de uma pergunta em língua natural denominam-se Sistemas de Pergunta-Resposta (SPR) e são o principal foco desta dissertação.

### 1.1 Motivação

Um SI clínico[44] tem como objetivo auxiliar o médico/utilizador na RI sobre um problema/situação clínico. A rapidez e facilidade com que o médico atinge/visualiza a informação pretendida é determinante para a qualidade do serviço prestado e, em urgência, para o estado clínico do paciente. Perante um problema de decisão, o médico formula uma pergunta na sua língua natural que terá de converter na linguagem do SI (palavras-chave, componentes gráficos) para obtenção de um subconjunto do conhecimento a partir do qual

seleciona/inferir uma solução para o problema. Esta abordagem requer uma gestão operacional/funcional de etapas e componentes do SI que consome tempo e gera distração do foco clínico. Em SPR a pergunta em língua natural é diretamente aplicada ao SI, originando um subconjunto do conhecimento orientado ao objetivo pretendido ou uma resposta concreta (eventualmente) em língua natural. Desta forma, encontrar informação no SI requer apenas conhecimento do domínio para formulação de perguntas relevantes, permitindo diminuir erros operacionais e facilitar a utilização/aprendizagem por novos utilizadores.

## 1.2 Objetivo

Nesta dissertação é proposto um sistema para RI numa ontologia utilizando frases interrogativas em língua natural portuguesa. Este objetivo consiste num SPR, sem geração de língua natural, aqui implementado com programação em lógica contextual, utilizando a implementação *GNU Prolog/CX*<sup>1</sup>, e aplicado a uma ontologia, em formato *RDF/OWL*<sup>2</sup>, de relatórios neurovasculares, sendo a sua (macro) arquitetura composta por:

- conversão da ontologia OWL em factos e regras Prolog, organizados/segmentados em unidades de contexto;
- análise morfológica, sintática e semântica da frase interrogativa para obtenção de interpretações linguísticas, especificadas em estruturas para representação de discurso;
- seleção de uma das interpretações e de factos/entidades da ontologia que lhe sejam relevantes, por validação de regras/restrições;
- composição da resposta com metodologia variável consoante a existência de paciente/indivíduo alvo na pergunta, por validação de padrões e/ou coincidência de prefixos entre os elementos selecionados e a totalidade da ontologia;

O SPR proposto recebe do utilizador uma pergunta em português para a qual retorna um conjunto de palavras da ontologia, organizadas de acordo com a sua estrutura original, relevantes para a resposta.

## 1.3 Conceitos envolvidos

A resolução do SPR proposto envolve etapas comuns em SPR, tal como a análise linguística, e específicas ao SPR desenvolvido, essencialmente devido ao uso de ontologias e programação em lógica contextual.

---

<sup>1</sup><http://www.di.uevora.pt/gprolog-cx/>

<sup>2</sup><http://www.w3.org/TR/owl-ref/>

**Ontologias** Ontologias[28] são modelos para organização de conhecimento que descrevem conceitos e instâncias destes com atributos e relações entre si, podendo ser aplicadas a domínios genéricos (a Web) ou específicos (medicina) sendo que qualquer ontologia específica pode ser englobada numa ontologia genérica. São particularmente úteis na *Semantic Web*<sup>3</sup> para dedução de conhecimento visto que englobam a terminologia e estrutura/fluxo lógica de um domínio num formato processável computacionalmente. No entanto, uma ontologia tem que ser consistente (não conter axiomas contraditórios) para possibilitar uma interpretação autónoma e fiável por agentes (sistemas computacionais autónomos). A verificação de consistência é (tipicamente) feita por *reasoners* (motores de regras/inferência) e é particularmente útil em ontologias geradas (semi) automaticamente[31]. Numa ontologia consistente os *reasoners* permitem também deduzir implicações lógicas que descrevem o encadeamento implícito e explícito entre objetos do conhecimento.

Visto que os especialistas na terminologia e conceitos de um domínio não são necessariamente fluentes em ontologias (e/ou informática) existem ferramentas de auxílio à modelação de conhecimento para uso local (tais como Protégé<sup>4</sup>, Loom/PowerLoom e Swoop[34]) ou remoto/distribuído (tais como WebODE[5], Ontolingua e pOWL) que incluem funcionalidades auxiliares tais como desenvolvimento colaborativo e controlo de versões. A visualização total ou parcial de informação numa ontologia pode ser feita através de árvores hierárquicas (conhecimento terminológico) e listas/tabelas de axiomas (factos do domínio) ou através de grafos (conceitos/instâncias como nós e atributos/relações como arestas) sendo que a facilidade de perceção do conhecimento varia consoante a granularidade da representação, expressividade das relações e clareza/simplicidade da terminologia aplicada.

**Programação em Lógica Contextual** A programação em lógica contextual é uma extensão aos elementos base da programação em lógica. O formalismo de programação em lógica mais utilizado/divulgado denomina-se Prolog e tem várias implementações (tais como SICStus, XSB e GNU Prolog<sup>5</sup>) sendo que todas definem um programa Prolog como um conjunto de instruções compostas por cabeçalho (nome e argumentos) e corpo (factos/condições que definem a veracidade do predicado) agrupadas em predicados que representam regras e factos. Todas as instruções com o mesmo nome (functor) e número de argumentos (aridade) definem o mesmo predicado (referido de forma abreviada como  $\langle \text{functor} \rangle / \langle \text{aridade} \rangle$ ). O predicado de uma regra contém pelo menos uma instrução com corpo enquanto que os predicados de factos apenas têm cabeçalho. Os factos Prolog descrevem uma árvore[14] cujos nós correspondem a argumentos do predicado.

A inclusão de contextos na programação em lógica[11, 1] permite modularidade no acesso e representação destes elementos base através de unidades Prolog[45, 46] que identificam os contextos de um programa, sendo acessíveis/manipuláveis com operadores próprios da implementação utilizada. Desta forma é possível organizar/segmentar o conhecimento descrito em factos Prolog.

---

<sup>3</sup><http://www.w3.org/2001/sw/>

<sup>4</sup><http://protege.stanford.edu/>

<sup>5</sup><http://www.gprolog.org>

## **1.4 Estrutura da dissertação**

O capítulo 2 descreve alguns aspetos comuns em SPR, para contextualização da implementação proposta no capítulo 3 que também inclui as decisões e tecnologias inerentes ao desenvolvimento deste SPR. A sua avaliação experimental é descrita no capítulo 4 tal como a discussão dos resultados obtidos. O capítulo 5 engloba as conclusões do trabalho desenvolvido e indica extensões para melhoramento das funcionalidades implementadas.

## Capítulo 2

---

# Sistemas de Pergunta e Resposta

*Estado da arte em Sistemas de Pergunta e Resposta descrevendo isoladamente componentes e sistemas.*

Os SPR descrevem uma aplicação de Inteligência Artificial (IA) que implica entendimento, representação e extração/dedução de conhecimento descrito em língua natural. Servem de base tecnológica ao teste formulado por Alan Turing[65, 58] que define um sistema/agente como inteligente quando, para as mesmas perguntas, um observador humano não consegue distinguir as suas respostas das geradas por um humano.

Enquanto sub-área da IA os SPR podem ser classificados consoante aplicam mecanismos de IA forte ou fraca[2] sendo que um SPR forte pode ser considerado independente do domínio e baseado em inferências genéricas enquanto que um SPR fraco é dependente de um (ou vários) domínio e utiliza componentes dependentes de trabalho humano (corpora anotados, léxicos) descrevendo uma inferência informada.

O desenvolvimento de um SPR envolve técnicas de PLN para entendimento de língua natural (na pergunta e/ou no repositório de conhecimento) e geração de estruturas linguísticas (para a resposta). O PLN permite extrair informação linguística numa frase por análise morfológica, sintática e semântica eventualmente complementada com Reconhecimento de Entidades Mencionadas (REM) para recolha de informação específica. Tipicamente o PLN em bases de conhecimento de um SPR é feito numa etapa de pré-processamento para redução do tempo de resposta.

A informação obtida da frase interrogativa submetida ao SPR é utilizada para restringir a base de conhecimento na RI relevante para a pesquisa. A organização do conhecimento no modelo de representação influencia o desempenho da RI visto que determina as capacidades de decisão/inferência e de expansão/adaptação.

Tanto os SPR como as suas componentes de PLN e RI derivam da IA, sendo áreas de estudo para técnicas/metodologias de implementação de agentes classificados como inteligentes por um observador humano[18]. Ainda que a definição/perceção de inteligência seja dependente do contexto/conceitos do observador, existem funcionalidades elementares aos SPR que são classificadas como inteligentes por qualquer observador tais como o entendi-

mento de língua natural por identificação de elementos linguísticos relevantes ou a recolha de elementos relevantes para resposta a uma pergunta.

Na secção 2.1 são descritos componentes/tecnologias envolvidos no desenvolvimento de um SPR enquanto que a secção 2.2 apresenta combinações típicas destes componentes e respetivas implementações.

## 2.1 Componentes

O conceito de SPR implica que o utilizador especifique a informação que pretende numa pergunta em língua natural e que a resposta seja uma seleção de elementos do conhecimento, representados/enumerados como ocorrem na base de conhecimento ou sob a forma de uma frase em língua natural. Estas etapas envolvem componentes de PLN e RI que podem ser distribuídos em aglomerados/*frameworks*[20, 16, 36] eventualmente orientados a SPR[59, 49, 3] pela inclusão de ferramentas previstas no âmbito de um SPR.

A variedade de componentes de PLN e RI sugere uma aplicação conjunta para melhor entendimento da língua natural e relevância de factos selecionados sendo que a complexidade e desempenho computacional do SPR aumenta proporcionalmente ao número de componentes utilizado. Em SPR complexos podem ser aplicados conceitos de sistemas distribuídos/paralelos eventualmente com controlo/seleção de componentes a utilizar através de técnicas de Aprendizagem[62].

### 2.1.1 Processamento de Língua Natural

Todo o componente que requer e/ou retorna apenas língua natural envolve PLN, logo os SPR são também componentes de PLN. Na década de 90 o desenvolvimento de componentes de PLN era restrito a ambientes de investigação[53, 10] e focado na língua inglesa sendo considerados essenciais para a disseminação dos SPR (e SI) no quotidiano humano devido à facilidade/naturalidade de comunicação entre Homem e máquina. Atualmente a maioria destes mecanismos (e respetivos recursos linguísticos) é considerada utilizável/estável na língua inglesa estando a investigação centrada noutras línguas[7] (particularmente europeias) incluindo o português[27, 24].

**Linguística** Em qualquer SPR, a pergunta submetida e/ou os textos que compõem a sua base de conhecimento requerem uma interpretação linguística para extração/adaptação dos seus constituintes a uma representação lógica/formal, interpretável pelos componentes de RI e de composição da resposta. A análise sintática da pergunta/frase permite isolar elementos distintos (tipo de pergunta, foco) e identificar funções/estruturas sintáticas podendo ser obtido maior detalhe lexical de cada elemento (género, número, categoria gramatical) através da sua morfologia.

Os elementos da análise sintática podem ser organizados numa estrutura, independente da pergunta/frase original, que representa a sua análise semântica, permitindo agrupar palavras correspondentes ao mesmo elemento e elementos com a mesma utilidade na definição do significado da frase.

Outro complemento comum na análise linguística é o REM, englobando a marcação de entidades relevantes (tais como nomes próprios e datas) para expansão do conhecimento local (sobre uma frase) e/ou global (sobre o domínio) com informação complementar/alternativa.

### 2.1.2 Representação de conhecimento

Em SPR independentes do domínio ou sobre domínios extensos (com diversos subdomínios) é obtida maior abrangência/cobertura com a utilização de múltiplas bases de conhecimento que podem ser descritas com distintas representações e contentores/elementos de informação, o que origina um aumento do número de técnicas de RI que podem estar implícitas no formalismo de representação do conhecimento.

**Teoria para representação de discurso** A organização semântica de constituintes de um conjunto de frases pode ser feita de acordo com a teoria para representação de discurso[35, 22] que define uma estrutura própria denominada DRS (de *discourse representation structure*) para representação unívoca de elementos em cada frase (através de variáveis) e suas implicações/relações (em condições sobre variáveis). Uma DRS pode englobar informação da análise sintática[55] e do domínio/aplicação[26], entre outros aspetos[61]. Os contentores elementares de uma DRS podem ser aumentados com novos grupos[39] ou com subgrupos[25] sendo que também a própria representação de DRS (como conjunto de contentores) pode ser alterada[54].

**GNU Prolog/CX** Esta implementação de Prolog com contextos[1] constitui uma extensão ao interpretador/compilador GNU Prolog com predicados para aplicação de conceitos da programação contextual em lógica. Nomeadamente define-se o contexto de um programa Prolog através da aplicação de uma anotação/diretiva (regra Prolog sem cabeçalho) no início do respetivo ficheiro com o predicado *unit/1* que associa o conjunto de predicados (factos e regras) no ficheiro ao argumento de *unit/1* (contexto dos predicados) numa noção equivalente à de pacote/classe em linguagens de programação imperativa, embora com a vantagem de manipulação dinâmica. Estão disponíveis diversos operadores para manipulação de contextos/*units* tal como *:>/2* que aumenta dinamicamente o contexto atual com predicados de um novo contexto (primeiro argumento) para execução de um golo Prolog (segundo argumento).

**RDF/OWL** A linguagem OWL para representação de ontologias estende a sintaxe RDF (baseada em XML) para descrever conceitos, instâncias, atributos e relações de uma ontologia sendo que na terminologia OWL estes componentes são referidos respetivamente como classes, indivíduos, *datatype properties* e *object properties* podendo ser aumentados com outros formalismos de representação (restrições, listas).

O formato OWL é classificado em espécies de acordo com a expressividade dos predicados nele aplicados sendo cada espécie definida por um subconjunto do total de predicados disponível. Uma ontologia OWL que utilize apenas predicados para definição elementar de componentes representa a espécie *OWL Lite* cujo conjunto de predicados associado está

contido na espécie *OWL DL* (subsequentemente mais expressiva) que contém predicados para descrição de paradigmas das Lógicas Descritivas (DL é acrónimo de *Description Logic*) e representa o formato OWL ideal para aplicação de mecanismos de raciocínio/inferência. A utilização de predicados não contidos em nenhum destes conjuntos classifica implicitamente a ontologia na espécie *OWL Full* que apesar de permitir representar um domínio de forma completa e pormenorizada apresenta dificuldades quando interpretada computacionalmente visto que não é garantida a consistência dos factos representados limitando a sua aplicabilidade a sistemas com capacidades de dedução.

Baseado numa lógica descritiva, o OWL DL permite a utilização de *reasoners* (tais como FaCT++<sup>1</sup>, Pellet<sup>2</sup> ou HermiT<sup>3</sup>) para dedução de conhecimento implícito (consequências lógicas) na representação de conhecimento, agindo como interpretadores de OWL. Além da inferência baseada em mecanismos de raciocínio é também possível obter informação contextual sobre indivíduos ou axiomas que apresentam certa propriedade/restricção percorrendo a ontologia com um *Visitor pattern*[51]. No entanto, o RDF/OWL tem associado o formalismo de interrogação SPARQL<sup>4</sup> que constitui a principal forma para consultas formais/parametrizáveis.

## 2.2 Arquiteturas

Os SPR podem ser classificados consoante aceitam perguntas factuais (o foco é um único facto), não factuais (descrevem um conjunto de factos combinados para descrição do objetivo) ou ambas. A arquitetura de um SPR combina técnicas para processamento de uma base de conhecimento, entendimento do significado da pergunta submetida, recolha de informação relacionada e geração de uma resposta. As fases essenciais a um SPR incluem:

- Processamento da pergunta: representação da informação sintática e/ou semântica da pergunta numa linguagem formal/computacional, identificação do tipo da pergunta (quem, o quê, onde) e da resposta esperada (localização, nome), tradução automática, identificação de conceitos/alternativas (sinónimos, hipónimos/hiperónimos);
- Recuperação de informação: extração de documentos/respostas, identificação de factos na base de conhecimento relacionados com a pergunta (por comparação de prefixos/palavras, anotações semânticas);
- Processamento da resposta: avaliação/classificação da informação recuperada para seleção de factos com maior probabilidade de correspondência com a informação pretendida na pergunta, composição da resposta a retornar ao utilizador;

---

<sup>1</sup><http://owl.man.ac.uk/factplusplus/>

<sup>2</sup><http://clarkparsia.com/pellet/>

<sup>3</sup><http://hermit-reasoner.com/>

<sup>4</sup><http://www.w3.org/TR/rdf-sparql-query/>

**Implementações** Os SPR atuais distinguem-se pela variedade de técnicas de PLN e RI utilizadas na sua implementação sendo o Watson[23] o mais popular e abrangente em técnicas utilizadas. Outros SPR são particularmente inovadores por utilizarem informação gerada por/para utilizadores[69, 17] (perguntas frequentes, grupos de discussão), pela sua aplicação em domínios específicos[6, 4], línguas europeias[13, 19, 64] ou simultaneamente em diversas línguas[48], por permitirem a intervenção/discurso do utilizador para desambiguar possíveis respostas[21, 57] e/ou por utilizarem técnicas tais como padrões linguísticos[15], reconstrução de frases[33] ou DRS[55].

A avaliação (semi) automática de SPR[29, 50, 56] é também uma prioridade atual devido à constante expansão destes sistemas, particularmente em múltiplas línguas[52, 42], sendo que manualmente existem encontros dedicados ao teste/avaliação conjunta a SPR e suas tarefas de PLN e RI[68].

## 2.3 Sumário

Neste capítulo foram enquadrados os SPR enquanto subárea da IA. Os seus componentes principais foram apresentados descrevendo técnicas tipicamente envolvidas e possíveis implementações/formalizações. Estes e outros componentes foram situados nas etapas típicas da arquitetura de um SPR, sendo também brevemente descritas implementações que aplicam tecnologias, recursos e/ou metodologias distintas.



## Capítulo 3

---

# CliniQA: um SPR na medicina

*Arquitetura do SPR desenvolvido com descrição de componentes envolvidos e suas adaptações.*

O SPR alvo desta dissertação permite utilizar língua natural portuguesa para interrogar conhecimento factual representado numa ontologia em formato OWL, sendo apelidado de CliniQA (acrónimo de *Clinical Question Answering*) para simbolizar a sua aplicação (e respetivo estudo/teste) à ontologia de relatórios médicos/clínicos descrita na subsecção 3.1.1. Embora desenhado e testado com base nesta ontologia, o SPR inerente ao CliniQA é genérico e aplicável a qualquer ontologia que utilize todos os contentores elementares de informação definidos pelo formato OWL (indivíduos, classes e propriedades).

O CliniQA recorre ao analisador sintático e semântico descrito na subsecção 3.1.2 para obter a representação Prolog da informação linguística na pergunta formulada pelo utilizador. Esta informação é utilizada para restringir a representação Prolog da ontologia, previamente obtida a partir do formato OWL com o conversor descrito na subsecção 3.1.3. A resposta é obtida por identificação de padrões entre a informação restringida e a informação (na ontologia) sobre o indivíduo alvo da pergunta (identificado com a funcionalidade de reconhecimento de nomes próprios do analisador sintático/semântico); caso não seja especificado um indivíduo alvo a pergunta é aplicada a todos os indivíduos contidos na informação restringida. Para ambos os casos estão definidas regras excecionais para quando as condições/padrões típicos não se verificam.

### 3.1 Componentes base

Todos os componentes do CliniQA foram desenvolvidos exclusivamente em ambiente académico como protótipos pelo que são pouco visíveis/populares no âmbito do PLN e/ou ontologias. Além de funcionalidades comuns a ferramentas mais populares apresentam também características tecnológicas ou funcionais que os distinguem e pelas quais são mais apropriados ao desenvolvimento de um SPR em Prolog aplicado a ontologias.

### 3.1.1 Ontologia de Relatórios Neurovasculares

O conhecimento utilizado pelo CliniQA provem de uma ontologia[12], desenvolvida no âmbito do projeto MEDON<sup>1</sup>, cujo domínio combina elementos clínicos (da prática da medicina tais como pacientes e equipamentos) e médicos (teorias/conceitos da Biologia e Medicina tais como artérias e “AVC”, utilizados/agrupados em elementos clínicos) extraídos de forma semi-automática a partir de relatórios sobre exames neurovasculares. Estes relatórios incluem informação técnica real e informação pessoal fictícia numa hierarquia de classes/conceitos que inclui:

- fenómenos médicos observados (estenose, ateromatose carotídea, etc);
- entidades anatómicas e suas classificações (artérias, permeabilidade, etc);
- elementos operacionais e humanos no contexto de um exame (técnico, aparelho, etc);

Estas classes agrupam indivíduos/instâncias de acordo com a categoria (conceito do domínio) a que pertencem sendo também utilizadas para caracterização visto que um indivíduo pode estar em várias classes tal como ilustrado na figura 3.1. Alguns níveis/classes da hierarquia definem relações/propriedades com outras classes (herdadas por descendentes) restringidas em cardinalidade (máximo) e existência (pelo menos um) na aplicação aos respetivos indivíduos. A informação concreta dos relatórios é representada em indivíduos caracterizados por (meta) atributos tais como:

- nome e identificador de termos médicos de acordo com o sistema UMLS<sup>2</sup>;
- marca e características do instrumento utilizado no exame;
- data, local e identificador do relatório (informação fictícia);
- nome, idade e identificador do paciente (informação fictícia);
- nome do médico e do técnico (informação fictícia);

Os indivíduos em subclasses de “Observação” e “Diagnóstico” representam os fenómenos médicos descritos nos relatórios e identificados pelos nomes destas subclasses. Os indivíduos de “Diagnóstico” podem estar relacionados com os de “Observação” através da propriedade “eh\_suportado\_por” tal como ilustrado na figura 3.1. Ambos podem estar relacionados com indivíduos em classes de entidades anatómicas onde ocorre o fenómeno. Estes indivíduos são o principal repositório de conhecimento médico sendo acessíveis através da classe “Conclusão” apenas associada a “Relatório” (superclasse de “Neurovascular”).

O acesso a detalhes de um paciente (através do nome) é feito em indivíduos de “Paciente” que (tal como os de “Conclusão”) apenas estão relacionados com os de “Neurovascular”. Existem nomes repetidos permitindo testar o CliniQA num cenário (provável) de informação incompleta/parcial em que apenas é registado o primeiro e último nome de um paciente (invalidando a utilização do nome como identificador unívoco).

<sup>1</sup>Financiado pela Fundação para a Ciência e Tecnologia (FCT), tendo o título *MEDON - Ontologias para a modelação de dados e procedimentos médicos* e a referência PTDC/EIA/80772/2006.

<sup>2</sup><http://www.nlm.nih.gov/research/umls/>

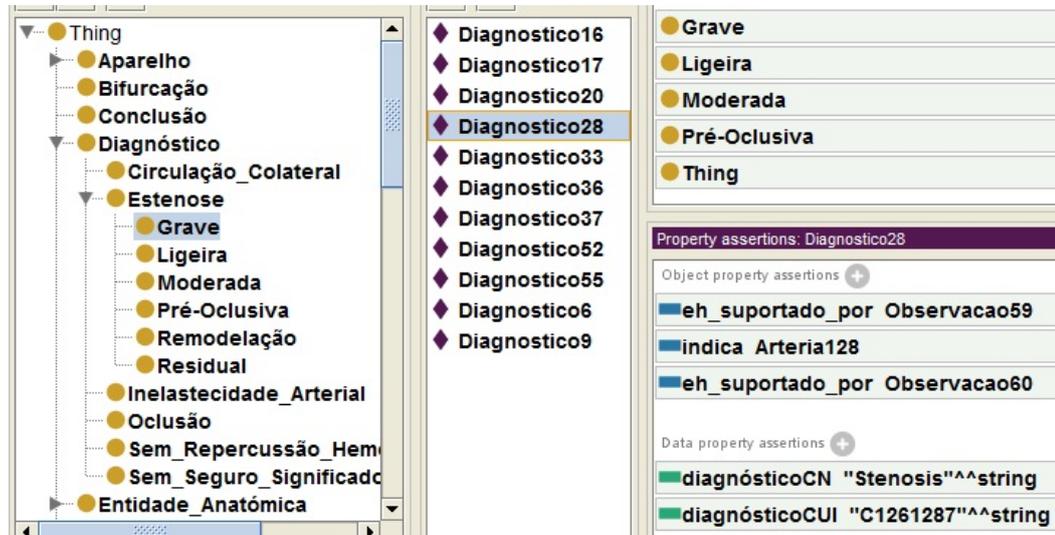


Figura 3.1: Representação no Protégé do indivíduo “Diagnóstico28” que pertence simultaneamente às classes “Ligeira”, “Moderada”, “Grave” e “Pré-Oclusiva”, está relacionado com “Observação” e “Arteria” e é descrito por informação do UMLS.

### 3.1.2 Analisador sintático e semântico com marcação de nomes próprios

O analisador sintático e semântico utilizado[39] foi desenvolvido especificamente para frases interrogativas em português e utiliza Prolog (implementação GNU Prolog) para representar os seus constituintes numa forma lógica/formal, organizados numa DRS e destacando nomes próprios mencionados. As DRS utilizadas definem um novo contentor de informação cujo único elemento/facto representa a ação/verbo principal da pergunta. O resultado do analisador é uma lista de possíveis interpretações da pergunta representadas pelo predicado *drs/3* (exemplificado na figura 3.2) que contem:

- referentes: lista de estruturas (com “-” como functor/separador) cujos constituintes identificam propriedades de sintagmas nominais (tais como número, tipo e advérbios complementares) e advérbios (apenas nome) sendo em ambos descrita uma variável que representa a respectiva entidade na DRS;
- condições: lista de factos que representam a atribuição de palavras na pergunta (functores) a variáveis, uma relação/dependência entre variáveis descrevendo um sintagma preposicional (predicado *rel/3*) ou um nome próprio detetado (predicado *nome/2*);
- verbo principal: lista com um facto que tem o infinitivo do verbo principal como functor e o tipo da pergunta (qual, quais, quem, etc) e sujeito a que se aplica (eventualmente numa estrutura com “:” com functor/separador que descreve o género do sujeito através do determinante que lhe está associado na pergunta) como argumentos;

```

1|      [ def-plu -[]-A-[], def-plu -[]-B-[], quais-plu -[]-C-[] ],
2|      [ nomes(A), pacientes(B), rel(de,A,B) ],
3|      [ diagnosticar(A,C) ]

```

Figura 3.2: Conjunto com a única DRS da pergunta «*Quais os nomes dos pacientes diagnosticados?*». Cada linha representa um constituinte da DRS e argumento do predicado *drs/3* (por ordem de ocorrência): 1- variáveis/referentes, 2 - factos/condições e 3 - ação/foco.

Os factos *nome/2* são gerados pelo marcador de nomes próprios tendo como argumentos a variável da entidade e o átomo/nome detetado. Para descrição de sintagmas preposicionais o predicado *rel/3* inclui como argumentos o átomo da preposição e as variáveis de entidades que a rodeiam na pergunta tal como ilustrado na figura 3.3.

O número de DRS varia consoante a complexidade sintática da pergunta relativamente ao número de entidades e conjugações possíveis entre estas, apesar de tipicamente apenas uma das DRS corresponder à pergunta formulada tal como ilustrado na figura 3.3. Algumas perguntas não geram DRS devido a estruturas sintáticas não previstas pelo analisador ou palavras inexistentes no dicionário utilizado internamente.

```

drs ([ def-sing -[]-B-[], def-sing -[]-C-[], que-plu -[]-D-[] ],
     [ aparelhos(D), exame(B), paciente(C), nome(C, 'Ana Maria ') ],
     [ utilizar(D,em:B,de:C) ])
| Significado: que utilizados em exame de paciente

drs ([ def-sing -[]-E-[], def-sing -[]-F-[], que-plu -[]-G-[] ],
     [ aparelhos(G), exame(E), paciente(F), nome(F, 'Ana Maria ') ], rel(de,E,F) ],
     [ utilizar(G,em:E) ])
| Significado: aparelhos utilizados em exame -> DRS mais apropriada

drs ([ que-H,[]-plu -[]-I-[], def-sing -[]-J-[], def-sing -[]-K-[] ],
     [ aparelhos(I), exame(J), paciente(K), nome(K, 'Ana Maria ') ], rel(de,J,K) ],
     [ utilizar(I,H,em:J) ])
| Significado: aparelhos que utilizados em exame

```

Figura 3.3: Múltiplas DRS da pergunta «*Que aparelhos foram utilizados no exame da paciente Ana Maria?*» ilustrando o significado inferível do verbo/ação principal por substituição das variáveis pelos respetivos funtores e conjugação do verbo.

Para correta interpretação, a pergunta submetida ao analisador (e a este SPR) não pode conter erros ortográficos visto que este utiliza a versão online do dicionário Priberam<sup>3</sup> como léxico global e a enciclopédia Wikipédia<sup>4</sup> como repositório de nomes próprios ambos compondo o seu dicionário local sempre que sejam utilizadas palavras (ainda) desconhecidas

<sup>3</sup><http://www.priberam.pt/dlpo/>

<sup>4</sup><http://pt.wikipedia.org>

nas perguntas formuladas. Este dicionário local é representado num ficheiro/repositório de factos Prolog (persistente entre atualizações) utilizado quando não há ligação à Internet.

### 3.1.3 XPTO Prolog Translation of Ontologies

O sistema XPTO[41] (acrónimo de *XPTO Prolog Translation of Ontologies*) permite representar uma ontologia RDF/OWL com programação em lógica contextual[11] (implementação GNU Prolog/CX[1]). Esta representação isola os componentes da ontologia em *units* (ficheiros Prolog exemplificados na figura 3.4) de acordo com a especificação do XPTO, nomeadamente:

- `class_X.pl`: classe da ontologia (identificada pelo XPTO com o número sequencial X) onde são descritas características próprias do OWL (*disjointWith/1*, *superClassOf/1*, etc) e restrições em propriedades da ontologia (*restriction/2* que inclui *constraint/2*);
- `prop_Y.pl`: propriedade da ontologia (identificada pelo XPTO com o número sequencial Y) onde são descritos predicados de acesso a todas as aplicações da propriedade (por *backtracking - item/0*), características próprias do OWL (*domain\_name/1*, *range/1*, etc), tipo (*object* ou *datatype - type/1*) e a sua inversa (*inverseOf/1*);
- `ontologies.pl`: mapeamento de nomes de classes (*class/4*) e propriedades (*prop/3*) na ontologia às respetivas *units* e descrição de ontologias envolvidas/importadas (*ns/3*);
- `individuals.pl`: agrupamento de indivíduos em classes (*individual\_class/2*) e triplos RDF da aplicação de propriedades (*property/3*) da ontologia ou do RDF/OWL que os relacionam com outros (*object properties*) ou com atributos (*datatype properties*); a ordem dos argumentos nestes predicados é relevante para a sua interpretação;

```

| ontologies.pl
class('http://input.owl', '', 'Conclusao', class_8).

| class_8.pl
axiom(disjointWith('Diagnostico')).
descriptor(restriction('obtida_atraves_de',
    constraint(someValuesFrom, 'Diagnostico'))).

| individuals.pl
individual_class('Conclusao9', 'Conclusao').
property('Conclusao1', descricao, 'Exames normais.').
property('Conclusao8', obtida_atraves_de, 'Diagnostico36').
property('Diagnostico36', rdf : type, 'Ligeira').

```

Figura 3.4: Excertos de *units* geradas com indicação dos respetivos ficheiros Prolog.

Estas *units* são geradas pelo XPTO através da execução do predicado *load\_ontology/1* no contexto da *unit* “dl” (própria do XPTO) onde são definidas as características/requisitos

para interpretação de ontologias OWL DL (único formato OWL implementado no XPTO). Os predicados nas *units* podem ser integrados em qualquer programa Prolog através de *consult/1* ao respetivo ficheiro sendo possível o acesso modular à informação (por *unit*) através das noções/operadores de contexto implementadas no GNU Prolog/CX[1].

## 3.2 Adaptações

Alguns componentes foram adaptados especificamente para utilização no CliniQA, nomeadamente a ontologia teve alterações de sintaxe/estrutura automatizadas numa aplicação Java e o dicionário utilizado pelo analisador foi aumentado manualmente (de acordo com problemas de reconhecimento identificados em perguntas de teste) permitindo interpretação de palavras/conceitos específicas ao domínio. O XPTO foi o único componente cujos recursos/estrutura internos não foram adaptados para utilização pelo CliniQA.

**Normalização da ontologia** Sendo descrita em língua portuguesa a ontologia utilizada no CliniQA contém símbolos para caracterização fonética/linguística (tais como acentos e cedilhas - diacríticos) cuja remoção permite que o seu processamento seja independente da Língua utilizada. A aplicação Java desenvolvida implementa esta funcionalidade separando os diacríticos da letra associada<sup>5</sup> para posterior identificação (com a expressão regular *InCombiningDiacriticalMarks*) e remoção. Sendo o OWL baseado em notação XML os caracteres com diacríticos estão codificados em ASCII sendo convertidos em *Unicode* para posterior remoção.

**Simplificação OWL** A sintaxe OWL utilizada na representação da ontologia não foi inicialmente aceite pelo XPTO pelo que foram necessários ajustes de sintaxe/estrutura e conteúdo automatizados pela aplicação Java desenvolvida com substituições (baseadas em expressões regulares) na representação textual da ontologia, nomeadamente:

- substituição de *owl:Thing* na definição de indivíduos pelo nome da primeira classe da ontologia a que o indivíduo pertence tal como descrito na figura 3.5;
- abreviatura de caminhos/URI<sup>6</sup> pela alteração de *rdf:about* para *rdf:ID*;
- remoção de anotações/comentários (predicado *rdf:comment*);
- remoção de “/” em nomes de elementos OWL;
- especificação do tipo de dados da *datatype property* “data” em todas as referências a esta («<data>» por «<data rdf:datatype="&xsd:date">»);
- introdução de conteúdo (irrelevante) em *datatype properties* cujo valor corresponde a uma sequência vazia de caracteres («<string"><» por «<string">NONE<»);

<sup>5</sup><http://download.oracle.com/javase/tutorial/i18n/text/normalizerapi.html>

<sup>6</sup><http://www.w3.org/TR/REC-rdf-syntax/>

- correcção da indentação/espacamento (por introdução de mudanças de linha);
- remoção de comentários XML;
- remoção de «<rdf:type rdf:resource="#owl:Thing"/>»;

```

| original
<owl:Thing rdf:about="#Arteria251">
  <rdf:type rdf:resource="#V4"/>
  <artériaCUI rdf:datatype="xsd:string">C0042559</artériaCUI>
  <artériaCN rdf:datatype="xsd:string">Structure of vertebral artery</artériaCN>
</owl:Thing>

| após normalização e substituições
<V4 rdf:ID="Arteria251">
  <arteriaCUI rdf:datatype="xsd:string">C0042559</arteriaCUI>
  <arteriaCN rdf:datatype="xsd:string">Structure of vertebral artery</arteriaCN>
</V4>

```

Figura 3.5: Exemplo de alterações para conformidade com o XPTO.

A representação obtida com estas alterações não é identificada<sup>7</sup> como OWL DL no entanto é aceite pelo XPTO. A última alteração não é necessária para que o XPTO interprete corretamente a ontologia, no entanto origina resultados distintos (igualmente válidos) em algumas das perguntas avaliadas tal como descrito no capítulo 4. A primeira e última alterações removem os indivíduos da classe genérica *owl:Thing* o que não altera o significado da ontologia visto que em teoria<sup>8</sup> todos os componentes pertencem a esta classe. O ficheiro OWL resultante destas alterações tem que ser codificado em UTF-8 com inclusão (no início) da *byte order mark* (BOM) correspondente a este sistema de codificação.

**Adenda do analisador** Alguns termos observados na ontologia não foram encontrados no dicionário da Priberam pelo que a análise de perguntas que os contém origina DRS menos corretas (se alguma). Para perguntas de teste utilizadas no desenvolvimento e avaliação do CliniQA apenas *ateromatose* (inexistente no Priberam) foi adicionada ao dicionário local como *nome(fem,sing,'ateromatose')* (nome comum feminino no singular). Entre outros termos inexistentes ou com classificação inadaptada ao domínio constam:

- nomes próprios: 2009 (ano de realização dos relatórios);
- nomes comuns femininos no singular: marca, permeabilidade;
- nomes comuns masculinos no singular: ecógrafo;
- adjetivos femininos no singular: colateral, ateromatosa, subclávia, carótida;

<sup>7</sup><http://www.mygrid.org.uk/OWL/Validator>

<sup>8</sup><http://www.w3.org/TR/owl-ref/>

- adjetivos femininos no plural: ateromatosas;

Apesar de não utilizadas em perguntas de teste estas palavras foram também adicionadas ao dicionário local pressupondo a sua classificação morfológica/gramatical por observação da sua utilização em documentos encontrados após pesquisa (por palavra) na Internet.

### 3.3 Arquitetura

A implementação do CliniQA descreve uma arquitetura modular em que predicados utilitários são agrupados/utilizados em predicados dedicados a PLN/SPR que definem a arquitetura global do sistema descrita na figura 3.6. Os predicados utilitários incluem pesquisa de texto em *individual\_class/2*, *property/3* ou *class/4* (por prefixo ou palavra completa), identificação de subclasses, extração de argumentos literais (valores de atributos), conversão para minúsculas, extração de prefixos com comprimento mínimo e coincidência de padrões textuais (com e sem normalização para minúsculas) entre outros.

A interface de língua natural descrita no diagrama da figura 3.6 representa o início de cada interação com o CliniQA (submissão de pergunta, obtenção de resposta) estando definidas componentes de pré-processamento anteriores a este ciclo não ilustradas no diagrama para orientação deste a tarefas específicas a PLN/SPR. Nomeadamente o CliniQA requer:

- localização do acesso principal ao analisador sintático/semântico (*main.pl*). Contem o predicado *analiseSemantica/2* que tem em argumento o caminho para um ficheiro apenas com a pergunta a analisar (gerado automaticamente na interface de língua natural) e uma lista/variável onde são registadas as DRS resultantes;
- localização da pasta que contem as *units* geradas pelo XPTO para a ontologia OWL. Caso este procedimento seja incorporado no SPR (numa versão mais flexível/adaptável do XPTO) apenas é necessário especificar a localização da ontologia;

As etapas até ao processamento das DRS são detalhadas na subsecção 3.3.1 enquanto que as etapas até à extração/identificação de informação (relativa a indivíduos) que dá origem à resposta são descritas na subsecção 3.3.2. A forma como a informação é recolhida da ontologia a partir do subconjunto inferido da pergunta é descrita na subsecção 3.3.3.

#### 3.3.1 Processamento da pergunta

A interface de língua natural do CliniQA ilustra a simplicidade desejável num SPR ao requerer apenas a introdução de uma frase interrogativa em língua natural para obtenção de informação previamente registada na ontologia.

**Seleção da melhor DRS** A pergunta formulada é submetida ao analisador que retorna uma lista de DRS de onde é selecionada a primeira que verifique uma das seguintes condições (enunciadas por ordem de verificação):

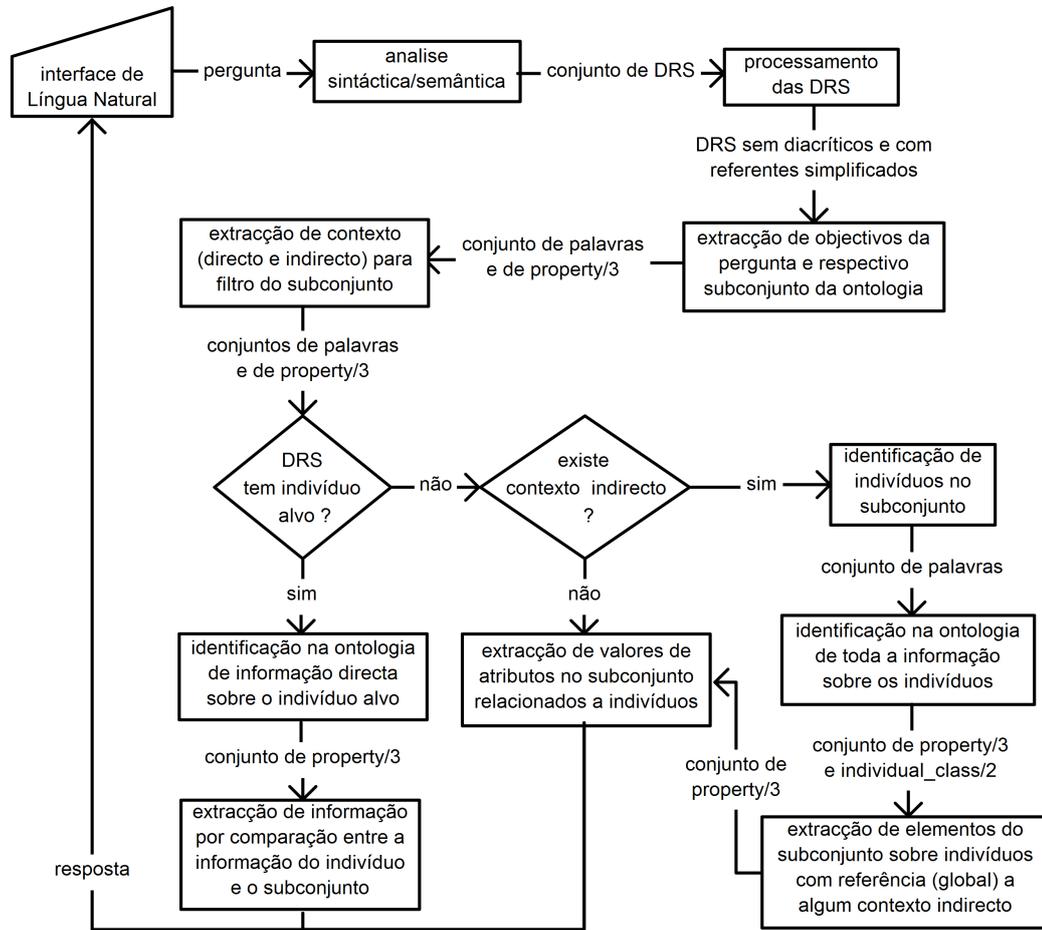


Figura 3.6: Arquitetura global do CliniQA.

- existem 2 *rel/3* em que a terceira variável de um é a segunda variável do outro, isto é  $rel(\_,\_,A)$  e  $rel(\_,A,\_)$  (a ordem de ocorrência é irrelevante);
- existe 1 *nome/2* cuja variável ocorre isoladamente noutra condição da DRS;

Estas condições foram inferidas manualmente através da observação de resultados do analisador; quando nenhuma se verifica é seleccionada a primeira DRS do conjunto.

**Normalização da DRS** Visto que o analisador sintático/semântico requer a utilização de ortografia correta a DRS resultante pode conter diacríticos. Estes são removidos numa etapa de normalização para agilizar futuras comparações entre a pergunta/DRS e a ontologia (normalizada numa fase de pré-processamento descrita na secção 3.2). Visto que o Prolog não dispõe de ferramentas (bem documentadas) para identificação destes marcadores foi utilizada uma aplicação auxiliar em Java que separa estas marcas da letra associada para

posterior identificação e remoção, tal como implementado para a normalização da ontologia. O Prolog comunica com esta aplicação através do sistema operativo, sendo utilizado um ficheiro de texto para transferir informação entre ambientes.

**Simplificação de referentes** Dos referentes obtidos pelo analisador são selecionados os que contém palavras da pergunta, sendo a sua estrutura simplificada de forma a conter apenas a respetiva variável e a palavra da pergunta. Pelo menos um dos referentes simplificados contem a palavra que define o tipo da pergunta, utilizado para RI específica a tipos de pergunta interpretáveis pelo CliniQA. Atualmente apenas são interpretados os tipos “onde”, “quando”, “qual”, “quais”, “que” e “quem” embora também sejam aceites perguntas com tipos não definidos na implementação. Caso a variável associada a um destes referentes seja partilhada com algum dos factos no conjunto de condições da DRS, o functor deste corresponde ao sujeito/objetivo da pergunta que fica assim isolado de outros objectivos inferidos.

### 3.3.2 Recuperação de informação

Para comparar a DRS e a ontologia é necessário selecionar informação relevante dispersa nestas bases de conhecimento. São definidas regras para inferência de objetivos e contexto a partir da DRS, sendo as palavras resultantes pesquisadas na representação da ontologia. O functor do facto nas condições da DRS que contem a variável do referente que descreve o tipo da pergunta é adicionado aos conjuntos de objetivos e contexto (se pertencente aos tipos de pergunta interpretáveis pelo CliniQA) para garantir a sua inclusão na pesquisa, embora tipicamente esteja incluído num destes conjuntos visto que as regras para inferência de objetivos e contexto englobam a maioria dos funtores na DRS.

**Objetivo** Da observação a DRS é possível inferir que os objetivos/foco da pergunta (o que se pretende saber) correspondem aos funtores de factos no conjunto de condições da DRS (exceto *rel/3*) que partilham variáveis com o facto do verbo principal. Estas variáveis podem ocorrer em múltiplos factos nas condições da DRS identificando um nome composto (de acordo com o analisador) tal como ilustrado na figura 3.7 para “informação clínica”.

```

- DRS normalizada :
[[[]-sing-[]-A-[],def-sing-[]-B-[],qual-sing-[]-C-[]]
 [ clinica(A), informacao(A), paciente(B), nome(B, Paula Rosado), rel(de,A,B)]
 [ ser(C,a:A) ]

-- referentes simplificados: [qual - C]
-- objetivo: [clinica, informacao]
-- contexto: [Paula Rosado, nome, paciente]

```

Figura 3.7: Informação recolhida automaticamente da pergunta “Qual a informação clínica da paciente Paula Rosado?”.

A pesquisa de objetivos na ontologia é feita no contexto da *unit* de indivíduos e apenas em factos *property/3*, gerando um subconjunto desta composto por *property/3* em que

ocorre (em argumento) um prefixo de alguma palavra no conjunto de objetivos. Este prefixo tem comprimento mínimo (tamanho do prefixo é maior que o do sufixo removido) permitindo a identificação de palavras na ontologia semelhantes às descritas na pergunta embora aumentando o ruído em prefixos comuns (tal como “nome” cujo prefixo é comum às propriedades *nomeMedico* e *nomePaciente*).

**Contexto** Os objetivos inferidos referem-se a um assunto/sujeito (sobre quem se pretende saber) aqui referido como contexto da pergunta. Este contexto corresponde a funtores de factos que contém a variável em *rel/3* (caso exista) não associada a um objetivo (contexto direto) sendo que quando esta se refere a um *nome/2* também o nome próprio (literal no último argumento) é recolhido, como ilustrado na figura 3.7. Este procedimento é aplicado recursivamente substituindo o conjunto (inicialmente) de objetivos por cada contexto inferido, de forma a permitir inferência com múltiplos *rel/3* (contexto indireto) em que algum não contém variáveis de objetivos.

O functor do verbo principal é também incluído (isoladamente) como contexto desde que composto por mais do que 3 caracteres de forma a evitar palavras irrelevantes (“ter” e “ser” são os mais comuns). Apesar do contexto ser descrito num único conjunto de palavras são gerados internamente subconjuntos deste consoante a inferência/proveniência (verbo principal, contexto directo/indirecto ou referentes) tal como ilustrado na figura 3.8. As palavras do contexto são pesquisadas no subconjunto obtido anteriormente (a partir dos objetivos) por ocorrência de prefixos (com comprimento mínimo como definido anteriormente) destas nos argumentos de *property/3*.

### 3.3.3 Composição da resposta

Com informação inferida da DRS e selecionada da ontologia é possível compor um aglomerado de informação que responde à pergunta formulada através da indicação/descrição de elementos relevantes na ontologia. A composição deste aglomerado envolve distintas fontes de informação e metodologias que variam consoante a existência de indivíduo/sujeito alvo na pergunta. No entanto, o subconjunto da ontologia (resultante da inferência de contextos) é comum entre estas metodologias visto que contem a informação, restringida da ontologia a partir da pergunta, mais relevante/apropriada para resposta. Sempre que possível (pelas regras definidas) a informação selecionada deste subconjunto é aumentada recorrendo à ontologia.

**Grafos** A principal fonte de conhecimento para inferência de respostas são os factos na *unit* de indivíduos que sendo binários (*individual\_class/2*) ou correspondentes com triplos RDF (*property/3*) podem ser representados num grafo cujos nós são elementos da ontologia (indivíduos, classes ou valores) interligados por arcos rotulados com o tipo de relação (functor em *individual\_class/2*) ou identificador de propriedade (segundo argumento em *property/3*) e direccionados de acordo com a ordem dos argumentos nos factos como ilustrado na figura 3.9.

A informação direta sobre um indivíduo/elemento corresponde ao conteúdo dos nós que lhe estão diretamente ligados (sem nós intermédios) enquanto que a informação global

```

- DRS com 0 rel/3:
[[]-sing-[]-A-[],que-sing-[]-B-[]]
[idade(B),paciente(A),nome(A,Paula Rosado)]
[ter(B,a:A)]

-- objetivo: [idade,nome,paciente]
-- contexto: [idade]
-- referentes: [idade]

- DRS com 1 rel/3:
[def-plu-[]-A-[],def-plu-[]-B-[],quais-plu-[]-C-[]]
[nomes(A),pacientes(B),rel(de,A,B)]
[diagnosticar(A,C)]

-- objetivo: [nomes]
-- contexto: [diagnosticar,pacientes]
-- verbo principal: [diagnosticar]
-- contexto direto: [pacientes]

- DRS com 2 rel/3 ambas contendo o objetivo:
[[]-sing-[]-A-[],def-sing-[]-B-[],def-sing-[]-C-[],qual-sing-[]-D-[]]
[conclusao(A),paciente(C),relatorio(B),nome(C,Vanessa Capucho),rel(de,A,B),rel(de,A,C)]
[ser(D,a:A)]

-- objetivo: [conclusao]
-- contexto: [Vanessa Capucho,nome,paciente,relatorio]
-- contexto direto: [Vanessa Capucho,nome,paciente,relatorio]

- DRS com 2 rel/3 em que apenas uma contem o objetivo:
[[]-sing-[]-A-[],def-plu-[]-B-[],def-plu-[]-C-[],quais-plu-[]-D-[]]
[ateromatose(A),carotidea(A),nomes(B),pacientes(C),rel(com,C,A),rel(de,B,C)]
[ser(D,B)]

-- objetivo: [nomes]
-- contexto: [ateromatose,carotidea,pacientes]
-- contexto direto: [pacientes]
-- contexto indireto: [ateromatose,carotidea]

```

Figura 3.8: Contextos recolhidos e respetiva proveniência das palavras inferidas (conjuntos vazios omitidos).

corresponde ao conteúdo dos nós em todos os caminhos do grafo que começam no nó do indivíduo/elemento e terminam num valor/literal. Apesar de não utilizada pelo CliniQA esta representação (com a representação Prolog correspondente) foi utilizada na inferência das respostas do CliniQA visto que permite uma visão mais explícita/abrangente das consequências lógicas implícitas na ontologia.

**Indivíduo alvo** Em perguntas com sujeito identificado por nome próprio é possível inferir o indivíduo/identificador a que pertence na ontologia por identificação de um prefixo (com comprimento mínimo como definido anteriormente) do argumento atómico/literal de *nome/2* em argumentos de *property/3* visto que estes factos descrevem triplos RDF onde o nome próprio só pode ocorrer no primeiro ou terceiro/último argumento (o segundo/intermédio descreve o identificador da propriedade que os relaciona). Nesta ontologia o nome próprio só é relacionado com identificadores de indivíduos logo o argumento de *property/3* no extremo oposto ao do nome próprio (apenas) corresponde ao identificador do indivíduo

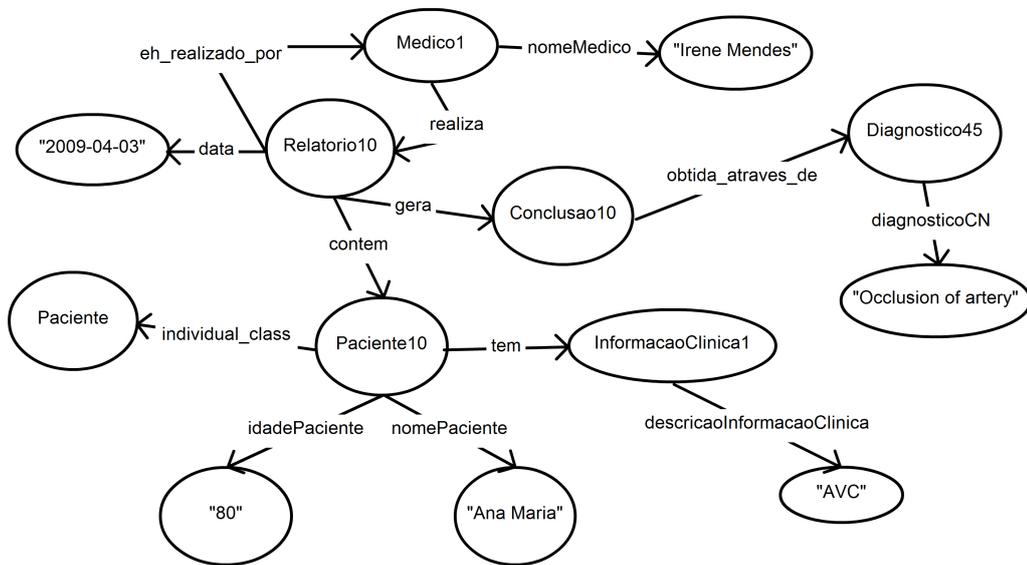


Figura 3.9: Excerto do grafo sobre a paciente “Ana Maria”.

alvo. Este é utilizado para extrair informação (da *unit* de indivíduos) diretamente relacionada com o indivíduo tal como a ilustrada na figura 3.10.

```

individual_class('Paciente10', 'Paciente').
property('Paciente10', numeroProcessoPaciente, '104').
property('Paciente10', idadePaciente, '80').
property('Paciente10', nomePaciente, 'Ana Maria').
property('Paciente10', tem, 'InformacaoClinica1').
property('Paciente10', 'encontra-se', 'Servico1').
property('Relatorio10', contem, 'Paciente10').

```

Figura 3.10: Informação direta sobre a paciente “Ana Maria”.

A utilização de elementos da ontologia como indivíduos alvo depende da sua representação visto que a inferência de respostas do CliniQA implica que o identificador do indivíduo alvo (além de ter associado um nome próprio) seja referido em *datatype properties* (para perguntas sobre o indivíduo) ou esteja diretamente relacionado com um indivíduo referido em *datatype properties* (para perguntas sobre algo relacionado com o indivíduo). Na ontologia utilizada apenas nomes de pacientes e médicos podem ser utilizados como indivíduos alvo (os respetivos identificadores existem em *datatype properties* e estão relacionados com indivíduos das classes “InformacaoClinica” e/ou “Relatorio”, ambos referidos em *datatype properties*) visto que técnicos e outros elementos da componentes clínica do domínio (tais como relatório e ecógrafo) apesar de terem associado um atributo/identificador literal apenas estão relacionados com elementos sem atributos literais.

**Padrões** Caso exista um indivíduo alvo na pergunta a metodologia a utilizar na recolha da informação (da ontologia) que compõe a resposta é determinada por padrões/regras baseados no índice de ocorrência do identificador do indivíduo (ou elemento diretamente relacionado com este) em argumentos de factos *property/3* no subconjunto da ontologia (inferido da pergunta) e na informação direta sobre o indivíduo (identificador existe em ambos). Apesar de inferidos por observação de coincidências entre factos os padrões podem ser enquadrados consoante a origem do conhecimento pretendido, nomeadamente:

- perguntas sobre o indivíduo alvo: identificador relacionado com o mesmo elemento no subconjunto e na informação do indivíduo (facto *property/3* comum);
- perguntas sobre algo que está relacionado com o indivíduo alvo: elemento relacionado com identificador na informação do indivíduo e com outro elemento no subconjunto, tal como *property(Relatorio10,contem,Paciente10)* (da informação sobre o indivíduo) e *property(Medico1,realiza,Relatorio10)* (do subconjunto);

No primeiro caso é selecionado o elemento comum (relacionado com o indivíduo) e no segundo caso o elemento distinto/único (exceto o indivíduo) entre os dois factos (um de cada conjunto) que aplicam o padrão. Em ambos os casos a metodologia para obtenção da resposta varia consoante o tipo de informação no elemento selecionado é:

- valor: o elemento é a resposta (facto que aplica o padrão é uma *datatype property*);
- identificador: são identificados os factos *property/3* que contém o elemento em algum dos argumentos sendo selecionado o que tem um valor no extremo oposto ao do elemento (procedimento inverso ao descrito na deteção do identificador do nome próprio) e retornado em formato simplificado (por exclusão de terminologia Prolog) como resposta (facto que aplica o padrão é uma *object property*);

**Padrão excepcional** Na inferência de respostas para um indivíduo alvo existe ainda um padrão (excepcional) para perguntas sobre algo que está relacionado com o indivíduo alvo em que o elemento selecionado é um identificador que não ocorre em *datatype properties*. A recolha de informação para este caso utiliza uma abordagem menos declarativa visto que infere conhecimento de forma iterativa e com um valor estático (definido por observação como sendo ideal para as perguntas testadas) a limitar o número de iterações.

Neste caso são extraídos (da ontologia) os factos *property/3* em que ocorre o elemento selecionado sendo recolhidos (para composição da resposta) os valores que ocorrem neste subconjunto da ontologia. Este procedimento é aplicado recursivamente a cada identificador no subconjunto sendo (por analogia com a representação em grafo) sucessivamente expandidos/ derivados nós descendentes do elemento selecionado. Para evitar excesso de informação existem apenas 3 derivações (níveis do grafo) a partir do elemento selecionado sendo que os identificadores nos subconjuntos partilham o 'contador' de derivações.

Este é o único procedimento que utiliza para inferência as classes dos indivíduos (obtidas pelos factos *individual\_class/2*) e informação contida nas *units* destas (atualmente apenas *subClassOf/1*). Assim, além dos valores são também recolhidas as classes de cada

identificador nos subconjuntos, as subclasses destas classes e as subclasses das subclasses (para maior detalhe) constituindo uma estrutura de informação tal como a descrita na figura 3.11. Nesta estrutura palavras relacionadas com o que se pretende na pergunta são representadas num aglomerado de listas de identificadores de classes ordenadas consoante a profundidade em que estas ocorrem na inferência seguidas de propriedades de indivíduos relacionados com o indivíduo alvo (no contexto descrito na pergunta) contidos em alguma destas classes. Desta forma a informação sobre o elemento selecionado é aumentada a partir da ontologia para uma estrutura de tamanho controlado retornada como resposta.

```
[ Hipoplasia ] - [ Observacao ] - [ Vertebral -[]- Arteria ]
C0243069-[Observacao94-observacaoCUI]
Hypoplasia-[Observacao94-observacaoCN]

[ Vertebral ] - [ Arteria , Vaso_Sanguineo ] - [ Direito -[Esquerdo]-Lado ]
C0042559-[Arteria209-arteriaCUI]
Structure of vertebral artery -[Arteria209-arteriaCN]
```

Figura 3.11: Resposta à pergunta “Qual a conclusão do relatório da paciente Vanessa Capucho?” (impressa de uma lista Prolog). O relatório correspondente contém a conclusão “Hipoplasia da Artéria Vertebral Esquerda/Direita”.

**Pesquisa global** Além de perguntas específicas a um indivíduo alvo o CliniQA permite também perguntas para consulta global de informação sobre componentes da ontologia (apenas pacientes e médicos pelas razões acima descritas para indivíduos alvo). A resposta a estas perguntas é um conjunto de factos *property/3* cuja abordagem de recolha varia consoante a informação obtida da pergunta sendo que nas perguntas atualmente aceites apenas são relevantes os contextos direto e indireto. Apesar de ainda limitadas na aplicação a componentes da ontologia as perguntas sem indivíduo alvo podem ser enquadradas em:

- consulta genérica: não existe contexto indireto (tal como em “Quais os nomes dos pacientes diagnosticados?”);
- consulta restringida: existe contexto indirecto (tal como em “Quais os nomes dos pacientes com aterosclerose carotídea?”);

Na consulta restringida o contexto indireto corresponde à restrição imposta sendo comparado com a informação global de cada identificador no subconjunto da ontologia para seleção dos que verificam a restrição. A informação global corresponde a todas as palavras/elementos derivadas do identificador num processo semelhante ao utilizado para os padrões excecionais embora sem limite na profundidade da pesquisa (evitando ciclos, pesquisa termina com a expansão de todos os nós descendentes do identificador) e apenas para factos *property/3*.

Caso alguma das palavras no contexto indireto ocorra como prefixo (com comprimento mínimo como definido anteriormente) de algum dos argumentos de factos *property/3* obtidos na pesquisa global, o identificador que originou a pesquisa verifica a restrição e o respetivo facto (no subconjunto) é incluído no conjunto de factos que constitui a resposta. Em consultas genéricas a resposta é o subconjunto inferido da pergunta sendo apenas selecionados os factos *property/3* que contém valores (*datatype properties*) para permitir uma resposta legível/estendível pelo utilizador.

Perguntas sobre outros componentes ou restrições envolvem outras abordagens para obtenção do conjunto de factos que melhor descreve os atributos pretendidos para a categoria/classe de indivíduos alvo da pergunta. No entanto, qualquer conjunto inferido deve conter *datatype properties* visto que o CliniQA requer que pelo menos um dos factos no conjunto utilizado para composição da resposta contenha um valor literal/legível (independentemente da existência de indivíduo alvo). A interação planeada para o CliniQA pressupõe que de início seja feita uma pergunta de consulta para obtenção de nomes de pacientes diagnosticados a utilizar como indivíduos alvo.

### 3.4 Sumário

Neste capítulo foi descrito o SPR desenvolvido, incluindo a apresentação genérica de recursos e ferramentas utilizadas tal como as suas adaptações para concordância com o domínio clínico e formato da representação de conhecimento. Apesar de genérica, os exemplos/excertos que acompanham a descrição de ferramentas referem-se à aplicação destas ao domínio clínico. A arquitetura deste SPR foi descrita num grafo (visão global) e com detalhe de decisões e etapas, acompanhando a implementação do algoritmo correspondente desde a inicialização do SPR até à composição da resposta final. Foram também referidas propriedades e representações alternativas das ontologias que ajudam à compreensão da inferência, complementando os excertos da informação envolvida, ilustrados de forma faseada consoante a etapa que os origina.

## Capítulo 4

---

# Avaliação

*Avaliação do CliniQA simulando uma possível aplicação em ambiente real.*

O CliniQA foi testado relativamente à cobertura de perguntas do conjunto de teste descrito na secção 4.1 e qualidade do processo de seleção de factos que compõem cada resposta. As perguntas de teste foram geradas manualmente a partir de elementos/relações da ontologia e relatórios que a originaram, seguindo características organizacionais desta ontologia tal como o agrupamento do conhecimento em pacientes. A utilização de conceitos médicos em perguntas de teste foi validada em significado e aplicação (aos restantes elementos da pergunta) de acordo com vários exemplos e definições encontrados na Internet, apenas considerados quando ocorrem de forma semelhante/idêntica em mais do que um documento.

A remoção de *owl:Thing* como superclasse de todos os indivíduos da ontologia permite evitar a definição (na inferência de respostas) de uma regra de exclusão de factos associados a esta representação. Esta regra seria particularmente utilizada em perguntas cuja informação recuperada aplica o padrão excepcional visto que a sua resposta é uma composição de identificadores relativos ao objetivo da pergunta, tipicamente relacionado com *owl:Thing*.

### 4.1 Conjunto de teste

Foram formuladas 73 perguntas de teste obtidas (essencialmente) por observação do relacionamento entre elementos da ontologia, sendo que algumas descrevem o mesmo conteúdo/significado em distintos tipos de pergunta e/ou escrita. Todas as perguntas incluem pelo menos uma palavra utilizada na ontologia sendo que 25 ( $\approx 34\%$ ) têm indivíduo alvo (médico, paciente ou técnico). Foram excluídas perguntas com mais do que um nome próprio (tal como “O que há de comum entre a paciente Paula Rosado e o paciente Fernando Silva?”) visto que não está definida detecção de múltiplos indivíduos alvo. As adendas feitas ao dicionário do analisador sintáctico são aqui utilizadas para evitar falhas derivadas da (falta de) DRS. Entre as perguntas formuladas constam:

- “Quais os nomes dos pacientes ”: “com 80 anos de idade?”, “com estenose moderada e em que artérias?”, “com estenose moderada na artéria subclávia?”, “diagnosticados em Maio?”. Também formuladas para outras propriedades (idades) e em variantes que não referem a propriedade pretendida (prefixo “Quais os pacientes ”);
- “Quais as características comuns em pacientes com ”: “ateromatose carotídea?”, “estenose na artéria carótida?”, “placas ateromatosas calcificadas?”, “estenose na artéria carótida direita?”. Apenas foram utilizados os diagnósticos “estenose” e “ateromatose carotídea” visto que são representados de forma distinta na ontologia ilustrando distintas formas de recuperação de informação;
- consultas sobre todos os pacientes (com ou sem restrições): “Dos pacientes com artérias oclusas, em quais houve reabilitação de fluxo por circulação colateral?”, “Quais os diagnósticos que ocorreram em 2009?”, “Quantos pacientes foram diagnosticados com estenose?”, “Qual a média de idades dos pacientes?”, “Quais as conclusões para cada paciente?”, “Em que pacientes foi observada ateromatose carotídea?”;
- consultas sobre pacientes específicos: “Quando foi examinada a paciente Paula Rosado?”, “Qual a data do relatório da paciente Paula Rosado?”, “Quais as características das placas ateromatosas da paciente Paula Rosado?”, “Qual o ecógrafo utilizado no exame da paciente Paula Rosado?”, “Que aparelhos foram utilizados no exame da paciente Ana Maria?”, “Quais os segmentos arteriais permeáveis da paciente Paula Rosado?”, “Qual o diagnóstico da paciente Paula Rosado?”;
- consultas sobre pessoas no contexto de um relatório neurovascular: “Quais os médicos que realizaram relatórios?”, “Que exame foi executado pela técnica Sónia Mateus?”, “Quais os relatórios realizados pela médica Irene Mendes?”;
- consultas sobre informação clínica: “Quais os diagnósticos que ocorreram em 2009?”, “Em que artérias é comum ocorrer estenose?”, “Em que ano ocorreram mais casos de estenose?”;
- booleanas e de cálculo: “A paciente Paula Rosado tem ateromatose carotídea?”, “Qual o tipo de exame mais frequente?”;

Algumas das perguntas formuladas sugerem reduzida probabilidade de interpretação pelo analisador devido a estruturas sintáticas extensas ou complexas, ou pelo CliniQA devido a terminologia distinta da utilizada na ontologia ou tipos de resposta não implementados (booleanas e de cálculo). Existem 6 perguntas cujo tipo de resposta não está implementado na inferência do CliniQA, tendo sido incluídas devido à sua elevada probabilidade de ocorrência num SI clínico, particularmente para gestão e validação do conhecimento.

Existem também perguntas que a informação na ontologia apenas permite responder parcialmente, tal como em “Quais os pacientes com ateromatose carotídea e em que artérias?” em que existe informação sobre pacientes com ateromatose carotídea mas não sobre as artérias em que ocorre. Estas perguntas têm por objetivo testar a capacidade do SPR não responder.

Este conjunto de teste foi formulado tendo em conta futuros/possíveis desenvolvimentos no CliniQA, podendo ser útil em outros SPR sobre domínios médicos/clínicos. Para teste/manutenção das funcionalidades implementadas foram utilizadas:

- “Qual a data do relatório da paciente Paula Rosado?”
- “Qual a conclusão do relatório da paciente Vanessa Capucho?”
- “Quais os nomes dos pacientes diagnosticados?”
- “Quais as idades dos pacientes com ateromatose carotídea?”
- “Que idade tem a paciente Paula Rosado?”
- “Que informação clínica tem a paciente Paula Rosado?”
- “Onde se encontra a paciente Paula Rosado?”
- “Quem realizou o relatório da paciente Vanessa Capucho?”

Este subconjunto é utilizado após novos desenvolvimentos para garantir que funcionalidades anteriormente validadas continuam a ser corretamente aplicadas, representando todos os tipos de pergunta e metodologias de resposta definidas no CliniQA. As respostas às perguntas de teste não estão definidas sendo obtidas manualmente por observação da ontologia e relatórios que a geraram.

## 4.2 Métricas experimentais

Para obtenção de valores quantitativos do desempenho do CliniQA foram medidas precisão, acerto e cobertura relativamente a resultados/respostas finais (macro avaliação) e a factos recuperados da ontologia (micro avaliação); e medida  $F^1$ , comum entre avaliações e ilustrada na figura 4.1. Estas métricas têm por foco as respostas do CliniQA e os mecanismos que as originam, logo não permitem/pretendem avaliar componentes intermédios tais como a RI da ontologia ou a interpretação de resultados do analisador linguístico.

$$\text{medida } F = \frac{2 \cdot \text{precisão} \cdot \text{cobertura}}{\text{precisão} + \text{cobertura}}$$

Figura 4.1: Definição de medida F, comum entre avaliações.

---

<sup>1</sup>É atribuída relevância idêntica à precisão e cobertura, logo é omitido o valor  $\beta$ [32] (igual a 1).

### 4.2.1 Seleção de factos

Avaliar o processo de composição de uma resposta implica contar, de entre os factos recuperados, aqueles que ocorrem e os que deveriam ocorrer na resposta final (se alguma). Estas contagens são agrupadas em medidas tipicamente utilizadas na avaliação de tarefas de aprendizagem/classificação[63], nomeadamente:

- verdadeiro positivo: facto recuperado que, corretamente, existe na resposta final;
- falso positivo: facto recuperado que, erradamente, existe na resposta final;
- verdadeiro negativo: facto recuperado que, corretamente, não existe na resposta final;
- falso negativo: facto recuperado que, erradamente, não existe na resposta final;

**Metodologia** A contagem destas medidas envolveu algumas regras, aplicadas a todas as perguntas, de forma a evitar contagens repetidas e sem significado, nomeadamente:

- A contagem de factos é feita subjetivamente visto que algumas respostas (tal como “Quais as conclusões para cada paciente?”) não estão totalmente expostas na ontologia, logo é considerado o maior número de factos recuperados que permite/impede o entendimento da resposta;
- As perguntas que não é possível responder diretamente com a ontologia (booleanas e calculadas - 6 no total) não foram consideradas;
- Em perguntas factuais é suficiente a existência do objetivo da pergunta em algum facto e não são avaliados/interpretados quantificadores. Assim, a resposta esperada à pergunta “Em que artérias é mais comum ocorrer estenose?” é composta por factos que contém o identificador/descrição de artérias associadas a algum diagnóstico/sub-classe de estenose;
- Em perguntas não factuais ou sobre informação que não está diretamente na ontologia (resposta descritiva) podem ocorrer factos na resposta que derivam da extensão/aumento de informação a partir de factos recuperados, não contados visto que não derivam (diretamente) da pergunta;
- As conjunções são interpretadas como separadores de condições que devem ocorrer na resposta, assim em “Quais os pacientes com ateromatose carotídea e em que artérias?” são contadas as artérias em que ocorre ateromatose carotídea ainda que não sejam referidos os pacientes (e vice versa);
- Quando não é referido na pergunta o que se pretende sobre certa entidade, é selecionado o conjunto mínimo de informação para identificação unívoca da entidade, assim a resposta a “Quais os pacientes com estenose moderada?” requer 2 factos por paciente: número de identificação e nome;

Estas contagens excluem duplicados (por pergunta) e factos sobre nomes próprios (recuperados separadamente na inferência) sendo utilizadas para definir as métricas de micro avaliação ilustradas na figura 4.2, à semelhança de outros SPR[60].

$$\text{precisão} = \frac{\text{verdadeiros positivos}}{\text{verdadeiros positivos} + \text{falsos positivos}}$$

$$\text{acerto} = \frac{\text{verdadeiros positivos} + \text{verdadeiros negativos}}{\text{verdadeiros positivos} + \text{falsos positivos} + \text{falsos negativos} + \text{verdadeiros negativos}}$$

$$\text{cobertura} = \frac{\text{verdadeiros positivos}}{\text{verdadeiros positivos} + \text{falsos negativos}}$$

Figura 4.2: Definição de cobertura, precisão e acerto para micro avaliação.

Estas métricas definem a qualidade do processo de seleção/classificação de factos na composição de uma resposta. Não foi avaliada a qualidade da recuperação dos factos, o que implicaria comparar os factos da ontologia com o seu subconjunto, recuperado a partir da pergunta.

#### 4.2.2 Respostas finais

A macro avaliação a resultados finais, observáveis pelo utilizador, é definida pelas fórmulas ilustradas na figura 4.3.

$$\text{precisão} = \frac{\text{número de perguntas corretamente respondidas}}{\text{número de perguntas respondidas}}$$

$$\text{acerto} = \frac{\text{número de perguntas corretamente respondidas}}{\text{número de perguntas submetidas}}$$

$$\text{cobertura} = \frac{\text{número de perguntas respondidas}}{\text{número de perguntas submetidas}}$$

Figura 4.3: Definição de cobertura, precisão e acerto para macro avaliação.

Nestas métricas, uma resposta é considerada correta se contém algum dos factos da ontologia referentes à informação perguntada, sendo irrelevante o detalhe com que são apresentados ou a completude da resposta. No entanto, uma resposta com (subjetivamente) demasiada informação irrelevante é considerada errada. Logo, esta avaliação constitui uma forma simples de analisar a perspetiva do utilizador pouco exigente.

**Analizador linguístico** O desempenho do CliniQA também depende do desempenho do analisador sintático/semântico visto que o CliniQA necessita da informação linguística para inferir uma resposta. Para medir o desempenho isolado do CliniQA, sem influência de falhas do analisador, foram geradas as fórmulas (alternativas) na figura 4.4, apenas para acerto e cobertura visto que as perguntas sem DRS apenas estão englobadas no denominador destas medidas.

$$\text{acerto} = \frac{\text{número de perguntas corretamente respondidas}}{\text{número de perguntas com DRS}}$$

$$\text{cobertura} = \frac{\text{número de perguntas respondidas}}{\text{número de perguntas com DRS}}$$

Figura 4.4: Definição de acerto e cobertura excluindo a influência do analisador linguístico.

**Indivíduos alvo** Além das medições globais (para todo o conjunto de teste) foi também medido o desempenho do CliniQA em perguntas com e sem indivíduo alvo visto que a metodologia de resposta varia consoante esta característica. Desta forma é possível isolar o tipo de perguntas (baseado na existência de indivíduo alvo) a que o CliniQA está mais apto a responder. Estas medições aplicam as fórmulas ilustradas na figura 4.3 a 2 conjuntos de teste de tamanhos diferentes (resultantes da divisão do conjunto de teste original).

Os distintos cálculos envolvidos na macro avaliação complementam a sua análise simples, permitindo medir a influência de aspetos exteriores à inferência do CliniQA ou próprios do domínio clínico utilizado.

## 4.3 Resultados

O cálculo das medidas de desempenho para macro e micro avaliação resultou nos valores apresentados na tabela da figura 4.5 com 2 casas decimais e sem arredondamento. A macro avaliação tem resultados distintos consoante se aplica ao conjunto de teste total/global, a subconjuntos deste compostos apenas por perguntas com ou sem indivíduo alvo ou apenas a perguntas com DRS (excluindo falhas do analisador) enquanto que a micro avaliação está definida apenas para o total do conjunto de teste.

	precisão	acerto	cobertura	medida F
<b>Macro avaliação</b>				
global	0.64	0.21	0.34	0.44
global excluindo falhas do analisador	0.64	0.26	0.41	0.49
com indivíduo alvo	0.76	0.4	0.52	0.61
sem indivíduo alvo	0.45	0.1	0.22	0.29
<b>Micro avaliação</b>				
global	0.06	0.81	0.41	0.1

Figura 4.5: Valores de precisão, acerto, cobertura e medida F para macro e micro avaliação.

Os valores envolvidos no cálculo da medida F utilizam os valores de precisão e cobertura apenas com as 2 casas decimais ilustradas e não são representados para simplificação da tabela. A precisão não se altera quando são excluídas as falhas do analisador visto que o seu cálculo apenas envolve perguntas com resposta.

#### 4.3.1 Discussão

Na macro avaliação, devido à elevada diversidade de perguntas e reduzida extensão do conjunto de teste formulado, o acerto e a cobertura foram as medidas com pior desempenho (ambas baseadas no total do conjunto de teste) sendo a precisão a melhor classificada, particularmente no subconjunto de teste composto por perguntas com indivíduo alvo. Logo, as regras definidas são apropriadas para um número limitado de tipos/categorias de pergunta sendo pouco flexíveis em DRS mais extensas/complexas ou compostas por palavras pouco relacionadas com o vocabulário utilizado na ontologia.

O melhor desempenho é obtido em perguntas com um indivíduo alvo devido à forma como esta ontologia representa pacientes e médicos (permitindo gerar conjuntos de comparação aptos à aplicação dos padrões) e pelo bom desempenho do analisador sintático/semântico no reconhecimento de nomes próprios.

Na micro avaliação a precisão foi praticamente nula, influenciando o valor da medida F, devido a 5 perguntas para consulta de entidades com validação de restrições. Estas perguntas têm resposta final e incluem a palavra "artéria" na entidade ou na restrição, sendo o prefixo desta (tal como calculado no CliniQA) muito frequente em identificadores de indivíduos na ontologia, logo envolvendo muitos factos recuperados. No entanto, o mecanismo de pesquisa global para consulta restringida não é ativado devido a uma falha na deteção

do contexto indireto (necessário para esta inferência). Assim, o CliniQA classifica estas perguntas como sendo de consulta genérica, logo a resposta é composta por muitos dos factos recuperados gerando um elevado número de falsos positivos. Excluindo as contagens de verdadeiros e falsos positivos (únicos intervenientes no cálculo da precisão) destas 5 perguntas obtém-se uma precisão de 0.84 e uma medida F de 0.55.

**Redundância** No conjunto de perguntas utilizado para testes sistemáticos ao CliniQA é garantida a obtenção de uma resposta correta visto que o desenvolvimento foi baseado/sustentado nestas perguntas. No entanto, algumas respostas (consideradas corretas) incluem informação redundante tal como a ilustrada na figura 4.6 em que a resposta (descrita no relatório) original é composta apenas por “Ateromatose Carotídea, sem repercussão hemodinâmica” e “Tortuosidade na ACI Esquerda/ Direita, compatível com Estenose Ligeira/Moderada/Grave”.

```
[Ateromatose_Carotidea] - [Observacao] - [Sem_Repercussao_Hemodinamica-[]-Diagnostico]
C0577631-[Observacao21-observacaoCUI]
Carotid Atherosclerosis -[Observacao21-observacaoCN]

[ Interna ] - [ Arteria ,Carotida_ou_Carotida_Comum] - [ Direito -[Esquerdo]-Lado]
C0007276-[Arteria50-arteriaCUI]
Internal carotid artery structure -[Arteria50-arteriaCN]

[Ligeira ,Moderada,Grave] - [Diagnostico ,Estenose] - [Tortuosidade-[]-Observacao]
C1261287-[Diagnostico9-diagnosticoCUI]
Stenosis -[Diagnostico9-diagnosticoCN]

[Sem_Repercussao_Hemodinamica] - [Diagnostico] - [Ateromatose_Carotidea-[]-Observacao]
NONE-[Diagnostico8-diagnosticoCN]
NONE-[Diagnostico8-diagnosticoCUI]

[Tortuosidade] - [Observacao] -
- [ Interna -[]-Carotida_ou_Carotida_Comum,Conclusao-[]-Carotida_ou_Carotida_Comum]
C0333076-[Observacao22-observacaoCUI]
Tortuosity -[Observacao22-observacaoCN]

[Tortuosidade] - [Observacao] -
- [ Interna -[]-Carotida_ou_Carotida_Comum,Grave-[Ligeira,Moderada]-Estenose]
C0333076-[Observacao22-observacaoCUI]
Tortuosity -[Observacao22-observacaoCN]
```

Figura 4.6: Resposta à pergunta “Qual a conclusão do relatório da paciente Paula Rosado?”.

Este excesso/redundância de informação é característico do padrão excecional (descrito na secção 3.3.3) onde a RI abrange todos os arcos de um conjunto de nós do grafo da ontologia (profundidade máxima da pesquisa) que podem estar interligados gerando informação redundante. É particularmente notório em indivíduos alvo com elevada densidade relacional no grafo (tal como a paciente Paula Rosado) derivada da quantidade de conceitos/nós utilizados (e mapeados) no relatório original.

Neste tipo de resposta são representados todos os elementos relevantes sem uma ordem/seleção que permita um entendimento rápido/evidente da resposta (existem exceções

tal como o exemplo na figura 3.11) no entanto são incluídos identificadores que permitem uma pesquisa restrita na ontologia (descrevem o sub grafo relevante) para obtenção de informação adicional. Recorrendo a técnicas de PLN para ontologias[40] (uma aproximação à geração de língua natural) é possível obter uma resposta mais natural (e menos redundante) que facilite o entendimento humano.

## **4.4 Sumário**

Neste capítulo foi descrito o ambiente experimental utilizado para avaliação do SPR desenvolvido, nomeadamente a composição e particularidades do conjunto de perguntas para teste, as métricas para avaliação quantitativa, os resultados da aplicação destas e respetiva discussão. A descrição das métricas incluiu o seu significado e cobertura de funcionalidades deste SPR, tendo sido discutidos os resultados da sua aplicação relativamente a melhorias previsíveis/imediatas e razões de sucesso/insucesso.



## Capítulo 5

---

# Conclusão

*Descrição da utilização idealizada, valor subjetivo e trabalho futuro para a solução desenvolvida.*

O CliniQA demonstra uma possível interligação de programação em lógica contextual e ontologias para resolução de um SPR independente do domínio. Apesar do reduzido desempenho, ainda que com métricas simples, a tarefa de RI numa ontologia a partir de uma pergunta em língua natural foi bem sucedida, englobando vários tipos de sintaxe e semântica para perguntas formuladas em português e permitindo resposta a algumas perguntas não factuais, apesar de requerer um utilizador especializado no domínio para entendimento destas. Logo, o objetivo desta dissertação foi alcançado de forma satisfatória, visto que o desempenho da maioria das métricas ronda os 50%, contribuindo para o estado da arte dos SPR tal como descrito na secção 5.1. As características mais limitadoras para uso genérico do CliniQA são descritas na secção 5.2 podendo ser resolvidas, tal como o melhoramento do desempenho, com algumas das soluções/sugestões propostas na secção 5.3.

### 5.1 Contributo

O desenvolvimento do CliniQA originou um SPR inovador pela combinação de ontologias e Prolog para responder a perguntas formuladas em língua natural. Os componentes que o constituem são igualmente inovadores pela sua implementação (essencialmente) declarativa e utilização de tecnologias invulgares noutros SPR tais como DRS e contextos.

O Prolog facilita inferências/comparações baseadas em padrões (inerentes à sua natureza lógica) e permite representar conhecimento de forma simples, semelhante aos triplos do formato RDF. A utilização de contextos permite converter a ontologia OWL em Prolog, equivalente ao utilizado para representar a interpretação da pergunta, para que a inferência da resposta seja feita numa única linguagem de programação permitindo uma implementação organizada e eficiente.

Desta forma uma ontologia OWL pode ser manipulada/questionada através de um sistema de inferência lógico, com mais e melhores capacidades computacionais relativamente a linguagens próprias do OWL (tal como o SPARQL). Apesar da tradução de OWL para

Prolog ser possível com versões base da linguagem[67], o Prolog contextual permite definir e aceder a distintos contentores de informação, de forma simples e nativa, para utilização das capacidades organizacionais do formato OWL.

O Prolog é também utilizado noutros SPR[38, 66, 55]) para representação de informação linguística e/ou inferência com padrões, embora sem a utilização de contextos. A metodologia de padrões utilizada não coincide com outros SPR baseados em ontologias OWL[60, 30, 9, 47, 37] visto que tipicamente a pergunta em língua natural é convertida para uma pergunta numa língua formal/computacional específica ao OWL.

No CliniQA, a inferência de uma resposta ocorre exclusivamente em ambiente Prolog podendo ser aplicada a qualquer base de conhecimento Prolog. No entanto, o acesso segmentado/conceptual ao conhecimento e a facilidade de adaptação/aceitação a outros domínios só é possível devido à utilização de ontologias/OWL.

## 5.2 Limitações

O CliniQA foi desenvolvido para utilização em locais de atividade médica calculada (clínicas/consultórios) ou imprevisível (hospitais/urgências) apesar de estar ainda muito limitado em aplicação visto que foi desenvolvido para responder a um conjunto de perguntas consideradas como essenciais (entre outras) a um SPR para o domínio médico descrito na ontologia utilizada. Estas perguntas não representam todas as alternativas (em significado, estrutura sintática e variedade lexical) descritas no conjunto de teste formulado que é também pouco abrangente visto que foi desenvolvido a partir da ontologia excluindo perguntas ambíguas (por utilização de sinónimos) e sobre estados clínicos não representados (pertencentes ao domínio global de patologias neurovasculares).

**Restrições de vocabulário** A correção de uma resposta do CliniQA depende da cobertura de informação da ontologia (relativa à pergunta) no subconjunto desta inferido da pergunta. Visto que este subconjunto é obtido por coincidência de prefixos entre a ontologia e a pergunta é essencial que ambas utilizem terminologia comum logo, a eficácia da resposta a algumas perguntas testadas provém da utilização de palavras que na ontologia estão associadas ao contexto ou objetivo da pergunta tal como em “Quem realizou o relatório da paciente Vanessa Capucho?” em que a utilização da palavra “realizou” (contexto da pergunta) é suficiente para responder à pergunta visto que os indivíduos da classe “Médico” (de onde é extraída a resposta) são relacionados com indivíduos da classe “Paciente” pelas propriedades “realiza” e “eh\_realizado\_por” (ambas contendo um prefixo de “realizou”).

## 5.3 Trabalho futuro

Durante o desenvolvimento do CliniQA foram identificadas (possíveis) melhorias, descartadas da implementação atual por estarem fora do âmbito e objetivos principais definidos para esta dissertação sendo extensões ao CliniQA (de funcionalidade ou correção de problemas) nomeadamente (por fase, componente ou problema em que se enquadra):

- objetivos: os componentes de um nome composto devem ser extraídos na ordem em que ocorrem na pergunta permitindo processar o nome composto como uma única palavra/expressão visto que o significado das constituintes pode diferir quando interpretadas isoladamente (tal como acontece atualmente quando “ateromatose carotídea” origina um subconjunto da ontologia que refere ateromatoses de outros tipos/artérias);
- contextos: devem ser excluídas palavras descritivas de funcionalidades do analisador tal como “nome” (já acontece para a palavra “rel”, functor do facto que representa relações entre componentes na pergunta);
- prefixos: o primeiro prefixo extraído de uma palavra (a utilizar para comparação com outra) deve ser a própria palavra de forma a reduzir sucessivamente as possibilidades de coincidência a partir do caso mais desejável (coincidência da palavra completa). Actualmente não há controlo sobre a ordem de remoção/geração de sufixos;
- dicionário: na fase de recuperação de informação a utilização de um dicionário (eventualmente especializado para o domínio) permite aumentar a cobertura de factos da ontologia devido ao reconhecimento de palavras não contidas na sua terminologia;
- correcção ortográfica simples: anteriormente à interpretação da pergunta pelo analisador deve ser corrigida ortografia simples (acentos, hífen, etc) com recurso a dicionários. O analisador implementa uma funcionalidade de pesquisa no dicionário Priberam que pode ser reutilizada/extendida para recolher sugestões fornecidas por este (tal como ilustrado na sua interface Web);
- sugestões: caso a correcção ortográfica implique alterações mais complexas (letras em falta/excesso ou mal ordenadas) podem ser utilizados algoritmos de recuperação de informação[43] baseados em estatística<sup>1</sup> ou padrões/distâncias (Levenshtein entre outras<sup>2</sup>). Os métodos baseados em estatística são particularmente aptos perante sugestões ambíguas visto que atribuem uma probabilidade a cada sugestão;

O conceito de *unit* pode também ser aplicado aos subconjuntos da ontologia (para restrição de informação específica à pergunta) através da escrita de cada facto inferido no ficheiro de uma *unit* temporária (eventualmente constituindo um repositório/cache de *units* para reutilização de subconjuntos em perguntas com componentes semelhantes) em vez da atual lista Prolog o que resulta numa melhor usabilidade/isolamento do código apesar de não alterar o desempenho da solução com as medidas definidas (eventualmente altera o tempo de processamento da resposta caso exista *cache*).

O reconhecimento de múltiplos indivíduos na pergunta é também importante para comparação entre informação clínica de pacientes que é uma possível utilização do domínio médico descrito na ontologia. Em perguntas sobre um indivíduo alvo cujo nome é comum a múltiplos indivíduos (como acontece na ontologia utilizada) a resposta deve incluir e agrupar informação recuperada para cada identificador associado ao nome.

<sup>1</sup><http://www.norvig.com/spell-correct.html>

<sup>2</sup><http://www.infoq.com/articles/lucene-did-you-mean>

Existem outras melhorias ainda sem sugestão (objetiva) de resolução tais como detecção de plurais e géneros para validação das respostas, alternativas para a pesquisa de profundidade infinita/total em perguntas (sem indivíduo alvo) de consulta restringida e resolução de perguntas booleanas. A extensão do CliniQA com algumas destas funcionalidades implica também aumentar o corpus (particularmente) com vocabulário mais variado para melhor representação da utilização natural da língua portuguesa.

## 5.4 Sumário

Esta tese propôs um SPR para pesquisas numa ontologia recorrendo apenas a língua natural, tendo sido aplicado e avaliado num domínio clínico em português. Apesar das limitações e melhorias sugeridas, foi obtido um ambiente base/estável, implementado com programação em lógica contextual, para interligação de componentes úteis no âmbito de um SPR. A combinação dos componentes aqui selecionados permitiu aumentar o estado da arte em SPR com uma alternativa essencialmente baseada em lógica e documentar características específicas da aplicação de SPR em ambientes médicos. A arquitetura formulada não está limitada apenas a resposta a perguntas em língua natural, podendo também ser adaptada para manipulação de ontologias em ambiente lógico. Este trabalho foi desenvolvido no âmbito do projeto MEDON - *Ontologias para a modelação de dados e procedimentos médicos* - identificado pela referência PTDC/EIA/80772/2006.

---

## Bibliografia

- [1] Salvador Abreu and Daniel Diaz. Objective: in minimum context. In *International Conference on Logic Programming (ICLP), Mumbai, India*, pages 128–147. MIT Press, 2003.
- [2] Mario Alai. Artificial intelligence, logic of discovery and scientific realism.
- [3] Pranav Anand, David Anderson, John D. Burger, John Griffith, Marc Light, Scott A. Mardis, and Alexander A. Morgan. Qanda and the catalyst architecture. In *TREC*, 2001.
- [4] Jurgen Angele, Eddie Monch, Henrik Oppermann, Steffen Staab, and Dirk Wenke. Ontology-based query and answering in chemistry: Ontonova @ project halo. In Dieter Fensel, Katia P. Sycara, and John Mylopoulos, editors, *International Semantic Web Conference*, volume 2870 of *Lecture Notes in Computer Science*, pages 913–928. Springer, 2003.
- [5] Julio C. Arpírez, Oscar Corcho, Mariano Fernández-López, and Asunción Gómez-Pérez. Webode: a scalable workbench for ontological engineering. In *K-CAP '01: Proceedings of the 1st international conference on Knowledge capture*, pages 6–13, New York, NY, USA, 2001. ACM.
- [6] Sofia J Athenikos and Hyoil Han. Biomedical question answering: a survey. *Computer Methods and Programs in Biomedicine*, 99(1):1–24, 2010.
- [7] Jordi Atserias, Bernardino Casas, Elisabet Comelles, Meritxell González, Lluís Padró, and Muntsa Padró. FreeLing 1.3: Syntactic and semantic services in an open-source NLP library . In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, Genoa, Italy, 2006. ELRA.
- [8] Cory Barr, Rosie Jones, and Moira Regelson. The Linguistic Structure of English Web-Search Queries. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, October 2008.

- 
- [9] Roberto Basili, Dorte H. Hansen, Patrizia Paggio, Maria Teresa Pazienza, and Fabio Massimo Zanzotto. Ontological resources and question answering. In Sanda Harabagiu and Finley Lacatusu, editors, *HLT-NAACL 2004: Workshop on Pragmatics of Question Answering*, pages 78–84, Boston, Massachusetts, USA, May 2 - May 7 2004. Association for Computational Linguistics.
- [10] M. Bates and R.M. Weischedel. *Challenges in natural language processing*. Studies in natural language processing. Cambridge University Press, 1993.
- [11] E. Bertino, B. Catania, and G.P. Zarri. *Intelligent database systems*. ACM Press books. Addison-Wesley, 2001.
- [12] Luís Borrego. Criação de uma ontologia e respectiva povoação a partir do processamento de relatórios médicos. Master’s thesis, Universidade de Évora, December 2010. Paulo Quaresma (superv.);.
- [13] Antonio Branco, Lino Rodrigues, Joao Silva, and Sara Silveira. Xisque: An online qa service for portuguese. In Antonio Teixeira, Vera de Lima, Luis de Oliveira, and Paulo Quaresma, editors, *Computational Processing of the Portuguese Language*, volume 5190 of *Lecture Notes in Computer Science*, pages 232–235. Springer Berlin / Heidelberg, 2008.
- [14] I. Bratko. *Prolog programming for artificial intelligence*. International computer science series. Addison Wesley, 2001.
- [15] Philipp Cimiano, Michael Erdmann, and Gunter Ladwig. Corpus-based pattern induction for a knowledge-based question answering approach. In *Proceedings of the International Conference on Semantic Computing*, pages 671–678, Washington, DC, USA, 2007. IEEE Computer Society.
- [16] P Clarkson and R Rosenfeld. *Statistical Language Modeling Using the CMU-Cambridge Toolkit*, volume 97, pages 2707–2710. Citeseer, 1997.
- [17] Gao Cong, Long Wang, Chin-Yew Lin, Young-In Song, and Yueheng Sun. Finding question-answer pairs from online forums. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR ’08*, pages 467–474, New York, NY, USA, 2008. ACM.
- [18] Ben Coppin. *Artificial Intelligence Illuminated*. Jones and Bartlett Publishers, Inc., USA, 2004.
- [19] Luís Fernando Costa and Luís Miguel Cabral. Answering portuguese questions. In *Proceedings of the 8th international conference on Computational Processing of the Portuguese Language, PROPOR ’08*, pages 228–231, Berlin, Heidelberg, 2008. Springer-Verlag.
- [20] Hamish Cunningham, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. *GATE: A framework and graphical development environment for robust NLP tools and*

## BIBLIOGRAFIA

---

- applications*, volume 54, pages 168–175. Association for Computational Linguistics, 2002.
- [21] Milan Agatonovic Danica Damljanovic and Hamish Cunningham. Identification of the question focus: Combining syntactic analysis and ontology-based lookup through the user interaction. In Bente Maegaard Joseph Mariani Jan Odjik Stelios Piperidis Mike Rosner Daniel Tapias Nicoletta Calzolari (Conference Chair), Khalid Choukri, editor, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may 2010. European Language Resources Association (ELRA).
- [22] Jan Van Eijck and Hans Kamp. Discourse Representation in Context. In Johan van Benthem and Alice ter Meulen, editors, *Handbook of Logic and Language*. Elsevier, 2010.
- [23] D Ferrucci, E Brown, J Chu-Carroll, J Fan, D Gondek, A A Kalyanpur, A Lally, J W Murdock, E Nyberg, and J Prager. Building watson : An overview of the deepqa project. *AI Magazine*, 31(3):59–79, 2010.
- [24] Cláudia Freitas, Paulo Rocha, and Eckhard Bick. Floresta sint&#225;(c)tica: Bigger, thicker and easier. In *Proceedings of the 8th international conference on Computational Processing of the Portuguese Language, PROPOR '08*, pages 216–219, Berlin, Heidelberg, 2008. Springer-Verlag.
- [25] Norbert E. Fuchs, Stefan Hoefler, Kaarel Kaljurand, Gerold Schneider, and Uta Schwertel. Extended Discourse Representation Structures in Attempto Controlled English. Technical Report ifi-2005.08, Department of Informatics, University of Zurich, Zurich, Switzerland, 2005.
- [26] Norbert E. Fuchs, Kaarel Kaljurand, and Tobias Kuhn. Discourse Representation Structures for ACE 6.6. Technical Report ifi-2010.0010, Department of Informatics, University of Zurich, Zurich, Switzerland, 2010.
- [27] Pablo Gamallo Otero and Isaac Gonzalez Lopez. A grammatical formalism based on patterns of part of speech tags. *International Journal of Corpus Linguistics*, 16(1):45–71, 2011.
- [28] Thomas Gruber. Ontology (computer science) - definition in encyclopedia of database systems. Springer-Verlag, 2009.
- [29] Martine Adda-Decker Guillaume Bernard, Sophie Rosset and Olivier Galibert. A question-answer distance measure to investigate qa system progress. In Bente Maegaard Joseph Mariani Jan Odjik Stelios Piperidis Mike Rosner Daniel Tapias Nicoletta Calzolari (Conference Chair), Khalid Choukri, editor, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may 2010. European Language Resources Association (ELRA).

- 
- [30] Qinglin Guo and Ming Zhang. Question answering system based on ontology. *2008 7th World Congress on Intelligent Control and Automation*, 5009:3347–3352, 2008.
- [31] Peter Haase and Johanna Völker. Ontology learning and reasoning - dealing with uncertainty and inconsistency. In *Proceedings of the Workshop on Uncertainty Reasoning for the Semantic Web (URSW)*, pages 45–55, 2005.
- [32] George Hripcsak and Adam S. Rothschild. Agreement, the f-measure, and reliability in information retrieval. *J. of Am. Medical Informatics Association*, 12(3):296–298, 2005.
- [33] Bernard Jacquemin. A derivational rephrasing experiment for question answering. In Bente Maegaard Joseph Mariani Jan Odjik Stelios Piperidis Mike Rosner Daniel Tapias Nicoletta Calzolari (Conference Chair), Khalid Choukri, editor, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may 2010. European Language Resources Association (ELRA).
- [34] Aditya Kalyanpur, Bijan Parsia, Evren Sirin, Bernardo C. Grau, and James Hendler. Swoop: A web ontology editing browser. *Web Semantics: Science, Services and Agents on the World Wide Web*, 4(2):144–153, June 2006.
- [35] H. Kamp and U. Reyle. *From discourse to logic: introduction to model theoretic semantics of natural language, formal logic and discourse representation theory*. Number pt. 2 in *Studies in linguistics and philosophy*. Kluwer Academic, 1993.
- [36] M. Konchady. *Building search applications: Lucene, LingPipe, and Gate*. Mustru Pub., 2008.
- [37] László Kovács, Norbert Fuhr, and Carlo Meghini, editors. *Research and Advanced Technology for Digital Libraries, 11th European Conference, ECDL 2007, Budapest, Hungary, September 16-21, 2007, Proceedings*, volume 4675 of *Lecture Notes in Computer Science*. Springer, 2007.
- [38] Adam Lally and Paul Fodor. Natural language processing with prolog in the ibm watson system. *Building*, 24(2):1–4, 2011.
- [39] João Laranjinho. Agente inteligente para análise sintáctica/semântica e marcação de nomes próprios. Master's thesis, Universidade de Évora, 2010. Irene Rodrigues (superv.);
- [40] Shao Fen Liang, Donia Scott, Robert Stevens, and Alan Rector. Unlocking medical ontologies for non-ontology experts. In *Proceedings of BioNLP 2011 Workshop, BioNLP '11*, pages 174–181, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.
- [41] Nuno Lopes, Cláudio Fernandes, and Salvador Abreu. Contextual logic programming for ontology representation and querying. In Axel Polleres, D. Pearce, Edna Ruckhaus,

## BIBLIOGRAFIA

---

- Gopal Gupta, and Stijn Heymans, editors, *Proceedings of the ICLP'07 Workshop on Applications of Logic Programming to the Web, Semantic Web and Semantic Web Services (ALPSWS 2007)*, volume 287, pages 27–42. CEUR Workshop Proceedings, October 2007.
- [42] Gilles Adda Brigitte Grau Dominique Laurent Véronique Moriceau Sophie Rosset Xavier Tannier Ludovic Quintard, Olivier Galibert and Anne Vilnat. Question answering on web data: The qa evaluation in quaero. In Bente Maegaard Joseph Mariani Jan Odjik Stelios Piperidis Mike Rosner Daniel Tapias Nicoletta Calzolari (Conference Chair), Khalid Choukri, editor, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may 2010. European Language Resources Association (ELRA).
- [43] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schtze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.
- [44] K. Mastrian. *Nursing Informatics and the Foundation of Knowledge*. Jones & Bartlett Publishers, Incorporated, 2011.
- [45] Paola Mello and Antonio Natali. Programs as collections of communicating prolog units. In Bernard Robinet and Reinhard Wilhelm, editors, *ESOP 86*, volume 213 of *Lecture Notes in Computer Science*, pages 274–288. Springer Berlin / Heidelberg, 1986.
- [46] Paola Mello and Antonio Natali. Objects as communicating prolog units. In *Proceedings of the European Conference on Object-Oriented Programming, ECOOP '87*, pages 181–191, London, UK, UK, 1987. Springer-Verlag.
- [47] A M Moussa and R F Abdel-Kader. Qasyo: A question answering system for yago ontology. *serscorg*, 4(2):99–112, 2011.
- [48] Gunter Neumann and Bogdan Sacaleanu. A cross-language question/answering-system for german and english. In Carol Peters, Julio Gonzalo, Martin Braschler, and Michael Kluck, editors, *Comparative Evaluation of Multilingual Information Access Systems*, volume 3237 of *Lecture Notes in Computer Science*, pages 101–109. Springer Berlin / Heidelberg, 2004.
- [49] Jun-Ping Ng and Min-Yen Kan. Qanus: An open-source question-answering platform. Technical report, 2010.
- [50] Djamel Mostefa Sophie Rosset Olivier Galibert Lori Lamel Jordi Turmo Pere R. Comas Paolo Rosso Davide Buscaldi Nicolas Moreau, Olivier Hamon and Khalid Choukri. Evaluation protocol and tools for question-answering on speech transcripts. In Bente Maegaard Joseph Mariani Jan Odjik Stelios Piperidis Mike Rosner Daniel Tapias Nicoletta Calzolari (Conference Chair), Khalid Choukri, editor, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may 2010. European Language Resources Association (ELRA).

- [51] Bruno C. Oliveira, Meng Wang, and Jeremy Gibbons. The VISITOR Pattern as a Reusable, Generic, Type-Safe Component. In *OOPSLA 2008 NASHVILLE, CONFERENCE PROCEEDINGS - MUSIC CITY USA, OOPSLA*, 1515 BROADWAY, NEW YORK, NY 10036-9998 USA, 2008. ACM SIGPLAN; SIGSOFT, ASSOC COMPUTING MACHINERY.
- [52] Bernardo Magnini Anselmo Peñas Álvaro Rodrigo Pamela Forner, Danilo Giampiccolo and Richard Sutcliffe. Evaluating multilingual question answering systems at clef. In Bente Maegaard Joseph Mariani Jan Odjik Stelios Piperidis Mike Rosner Daniel Tapias Nicoletta Calzolari (Conference Chair), Khalid Choukri, editor, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may 2010. European Language Resources Association (ELRA).
- [53] F.C.N. Pereira and B.J. Grosz. *Natural language processing*. Bradford Books. MIT Press, 1994.
- [54] Paul Piwek, Brigitte Krenn, Marc Schroeder, Martine Grice, Stefan Baumann, and Hannes Pirker. Rrl: A rich representation language for the description of agent behaviour in neca. In *Proceedings of AAMAS 2002 workshop: Embodied conversational agents*, Bologna, Italy, July 2002.
- [55] Paulo Quaresma and Irene Rodrigues. A logic programming based approach to qa@clef05 track. In *CLEF*, pages 351–360, 2006.
- [56] Silvia Quarteroni and Alessandro Moschitti. A comprehensive resource to evaluate complex open domain question answering. In Bente Maegaard Joseph Mariani Jan Odjik Stelios Piperidis Mike Rosner Daniel Tapias Nicoletta Calzolari (Conference Chair), Khalid Choukri, editor, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may 2010. European Language Resources Association (ELRA).
- [57] Luis Quintano and Irene Pimenta Rodrigues. Question/answering clarification dialogues. In *MICAI*, pages 155–164, 2008.
- [58] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach (2nd Edition)*. Prentice Hall, 2 edition, December 2002.
- [59] Nico Schlaefer, Petra Gieselman, and Guido Sautter. The ephyra qa system at trec 2006. In *In Proceedings of the Fifteenth Text REtrieval Conference (TREC)*, 2006.
- [60] Meltem Serhatli and Ferda N Alpaslan. An ontology based question answering system on software test document domain. *World Academy of Science Engineering and Technology*, 54:659–663, 2009.
- [61] C.S. Smith. *Modes of discourse: the local structure of texts*. Cambridge studies in linguistics. Cambridge University Press, 2003.

## BIBLIOGRAFIA

---

- [62] Daniel Sonntag. Distributed nlp and machine learning for question answering grid. In *Proceedings of the workshop on Semantic Intelligent Middleware for the Web and the Grid at ECAI*, 2004.
- [63] Håkan Sundblad. Question classification in question answering systems, 2007. Report code: LiU-Tek-Lic-2007:29.
- [64] Xavier Tannier and Véronique Moriceau. Fidji: Web question-answering at quaero 2009. In Bente Maegaard Joseph Mariani Jan Odjik Stelios Piperidis Mike Rosner Daniel Tapias Nicoletta Calzolari (Conference Chair), Khalid Choukri, editor, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may 2010. European Language Resources Association (ELRA).
- [65] A. M. Turing. *Computing machinery and intelligence*, pages 11–35. MIT Press, Cambridge, MA, USA, 1995.
- [66] Maria Vargas-vera and Enrico Motta. Aqua: A question answering system for heterogeneous sources.
- [67] Vangelis Vassiliadis, Jan Wielemaker, and Chris Mungall. *Processing OWL2 ontologies using Thea: An application of logic programming*, volume 2009. Citeseer.
- [68] Ellen Voorhees. The philosophy of information retrieval evaluation. In *In Proceedings of the The Second Workshop of the Cross-Language Evaluation Forum on Evaluation of Cross-Language Information Retrieval Systems*, pages 355–370. Springer-Verlag.
- [69] Guangyou Zhou, Li Cai, Jun Zhao, and Kang Liu. Phrase-based translation model for question retrieval in community question answer archives. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 653–662, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.



## Apêndice A

---

### Perguntas de teste

*Conjunto de teste com as 73 perguntas utilizadas para avaliação do SPR desenvolvido.*

Quais as conclusões para cada paciente?

Dos pacientes com artérias oclusas, em quais houve reabitação de fluxo por circulação colateral?

A paciente Paula Rosado tem ateromatose carotídea?

Em que ano ocorreram mais casos de estenose?

Em que artérias é comum ocorrer estenose?

Em que artérias é mais comum ocorrer estenose?

Em que artérias ocorre estenose?

Em que data foi realizado o relatório da paciente Paula Rosado?

Em que pacientes foi observada ateromatose carotídea?

Onde decorreu o exame da paciente Paula Rosado?

Onde foi examinada a paciente Paula Rosado?

Onde se encontra a paciente Paula Rosado?

Quais as artérias em que é mais comum a ocorrência de placas ateromatosas calcificadas?

Quais as artérias em que é mais comum a ocorrência de placas ateromatosas?

Quais as características comuns em pacientes com ateromatose carotídea?

- Quais as características comuns em pacientes com estenose na artéria carótida direita?
- Quais as características comuns em pacientes com estenose na artéria carótida?
- Quais as características comuns em pacientes com estenose?
- Quais as características comuns em pacientes com placas ateromatosas calcificadas?
- Quais as características comuns entre pacientes com ateromatose carotídea?
- Quais as características das placas ateromatosas da artéria carótida esquerda da paciente Paula Rosado?
- Quais as características das placas ateromatosas da paciente Paula Rosado?
- Quais as características que descrevem as placas ateromatosas?
- Quais as conclusões de diagnósticos a pacientes com algum tipo de estenose?
- Quais as conclusões de diagnósticos a pacientes com ateromatose carotídea?
- Quais as conclusões geradas?
- Quais as idades dos pacientes com ateromatose carotídea?
- Quais as idades dos pacientes diagnosticados?
- Quais os diagnósticos que ocorreram em 2009?
- Quais os médicos que realizaram relatórios?
- Quais os nomes dos médicos?
- Quais os nomes dos pacientes com 80 anos de idade?
- Quais os nomes dos pacientes com ateromatose carotídea?
- Quais os nomes dos pacientes com estenose moderada e em que artérias?
- Quais os nomes dos pacientes com estenose moderada na artéria subclávia?
- Quais os nomes dos pacientes com estenose moderada?
- Quais os nomes dos pacientes diagnosticados em Maio?
- Quais os nomes dos pacientes diagnosticados?
- Quais os pacientes com ateromatose carotídea e em que artérias?
- Quais os pacientes com ateromatose carotídea?
- Quais os pacientes com estenose moderada e em que artérias?

- Quais os pacientes com estenose moderada?
- Quais os pacientes diagnosticados em Maio?
- Quais os relatórios gerados?
- Quais os relatórios realizados pelo médico Irene Mendes?
- Quais os relatórios realizados?
- Quais os segmentos arteriais permeáveis da paciente Paula Rosado?
- Qual a conclusão do diagnóstico à paciente Paula Rosado?
- Qual a conclusão do relatório da paciente Paula Rosado?
- Qual a data do exame da paciente Paula Rosado?
- Qual a data do relatório da paciente Paula Rosado?
- Qual a idade da paciente Paula Rosado?
- Qual a informação clínica da paciente Paula Rosado?
- Qual a informação clínica de cada paciente?
- Qual a média de idades dos pacientes com artérias oclusas?
- Qual a média de idades dos pacientes?
- Qual o ano em que ocorreram mais casos de estenose?
- Qual o diagnóstico da paciente Paula Rosado?
- Qual o ecógrafo utilizado no exame da paciente Paula Rosado?
- Qual o ecógrafo utilizado?
- Qual o exame mais frequente?
- Qual o tipo de exame mais frequente?
- Quando foi examinada a paciente Paula Rosado?
- Quando foi realizado o relatório da paciente Paula Rosado?
- Quantos pacientes foram diagnosticados com estenose moderada?
- Quantos pacientes foram diagnosticados com estenose?
- Que aparelhos foram utilizados no exame da paciente Ana Maria?
- Que diagnósticos ocorreram entre Fevereiro e Maio de 2009?

Que exame foi executado pela técnica Sónia Mateus?

Que idade tem a paciente Paula Rosado?

Que informação clínica existe para a paciente Paula Rosado?

Que informação clínica tem a paciente Paula Rosado?

Quem realizou o relatório da paciente Vanessa Capucho?