# Um sistema de Pergunta-Resposta para uma base de Documentos

Paulo Quaresma
Irene Rodrigues
Universidade de Évora
Évora, Portugal
{pq,ipr}@di.uevora.pt

Carlos A. Prolo
PUCRS
Porto Alegre, Brasil
prolo@inf.pucrs.br

Renata Vieira
UNISINOS
São Leopoldo, Brasil
renata@exatas.unisinos.br

31 de Julho de 2005

## Resumo

Neste artigo apresentamos a metodologia seguida para a construção de um sistema de pergunta-resposta sobre uma base de documentos em Português. Descrevemos o sistema que tem dois módulos distintos: análise prévia dos documentos (extracção de informação) e processamento das perguntas (recuperação de informação). O nosso sistema procura fazer um processamento, do corpus e da perguntas, suportado em teorias da linguística computacional: análise sintáctica (gramática de restrições), seguida da análise semântica usando a teoria da representação do discurso e finalmente a interpretação semântica/pragmática usando ontologia e inferência lógica.

Apresentamos resultados da avaliação do seu desempenho que foi feita sobre dois conjuntos de documentos: textos de dois anos de jornais diários, o Público e a Folha de São Paulo (1994-1995); e um conjunto de documentos jurídicos: decisões do supremo Tribunal, tribunal da Relação e pareceres da Procuradoria Geral da República (cerca de 100.000 documentos).

# 1 Introdução

Este artigo descreve um projecto em desenvolvimento para construir um sistema de Pergunta-Resposta para o Português. Algumas das suas características são dependentes da Língua Portuguesa mas outras são independentes da Língua.

O sistema tem dois módulos: análise prévia dos documentos (extracção de informação) e processamento das perguntas (recuperação de informação).

O processamento, do corpus e da perguntas, é feito usando modelos e teorias actuais da linguística computacional. Este processamento inclui: a análise sintáctica das frases usando uma gramática de restrições, o analisador sintáctico Palavras [3]; a análise semântica usando a teoria da representação do discurso [5]; e finalmente a interpretação semântica/pragmática usando ontologia e inferência lógica.

O nosso sistema para representar a base de conhecimentos e a ontologia usa uma extensão ao Prolog, o ISCO[1, 2], que permite resolver alguns dos problemas que surgiram devido ao tamanho do corpus, como por exemplo: para dois anos do jornal "O Público" obtemos cerca de 9 milhões de entidades do discurso.

Os Sistemas de Pergunta-Resposta são um tópico actual na área de Processamento de Língua Natural. Algumas conferências internacionais têm iniciativas especiais para a avaliação de sistemas Pergunta-Resposta, por exemplo o TREC, Text REtrieval Conference (http://trec.nist.gov), e o CLEF, Cross Language Evaluation Forum (http://www.clef-campaign.org). Os sistemas avaliados nestas iniciativas são desenvolvidos para o domínio geral. No entanto também existem sistemas desenvolvidos para domínios específicos como por exemplo o domínio jurídico[6]. O domínio jurídico é uma área onde os sistemas de pergunta-resposta podem ser aplicados permitindo aos cidadãos uma melhor acesso à informação jurídica contida em documentos disponibilizados ao publico mas muitas vezes herméticos para um cidadão leigo.

O sistema de Pergunta-Resposta, para corresponder aos requisitos do Clef para avaliação, deve responde a uma pergunta em Língua Natural com base num conjunto de documentos. A resposta a uma pergunta é: um conjunto de palavras mais a identificação do documento e da frase de onde se extraiu a resposta. Por exemplo, para a seguinte pergunta:

Quem é a viúva de John Lennon?

O sistema desenvolvido retorna a resposta:

palavras: Yoko Ono - documento: publico/publico95/950807/001 - frase: 2

O sistema de pergunta resposta está vocacionado para responder a perguntas sobre:

- Lugares:

  Onde fica a Régua?

  Onde é que caiu um meteorito em 1908?

- Datas:

  Quando foi preso o Sr X?

  Quando morreu o Sr. X?

- Definições:

  O que é a ONU?

  O que é a Mouraria?

- Especificas:

  Quantas vezes foi o Sr X acusado de tráfico de droga?

  Quem foi preso por tráfico de droga?

  Que crimes cometeu o Sr Y?

No entanto, o sistema desenvolvido pode ser utilizado para desempenhar outras tarefas que não envolvam só o cálculo da "melhor" resposta a uma pergunta:

- Melhorar o desempenho de sistemas de recuperação de informação na tarefa de obter um conjunto de documentos relevantes para uma pergunta.

  Dada uma frase (pergunta ou não) em Língua Natural. O sistema de pergunta-resposta pode retornar todos os documentos que têm uma frase que pode verificar a pergunta.

  Para a pergunta "Quem é a viúva de John Lennon" o nosso sistema pode retornar todos os documentos que têm uma frase que torna verdadeiro o termo:

  $$\exists X, Y : viuva(X, Y), nome(X, ' John\_Lennon')$$

  Muitas vezes a reposta a uma pergunta não é única, pois a informação nos documentos não é precisa (textos de Jornais) e a forma como o módulo de extracção do sistema de Pergunta-Resposta processa a informação está sujeita à acumulação de erros (e.g. erros na análise sintáctica são transportados para as fases seguintes da extracção de informação).

- Auxiliar na tarefa de estruturar documentos de forma semi-automática.

  Por exemplo, no domínio Jurídico esta tarefa é muito importante, pois, em Portugal, a maioria dos documentos que estão digitalizados não têm qualquer estrutura associada. Para que possam ser introduzidos em bases de dados para pesquisa de informação o cálculo da estrutura dos documentos é uma tarefa essencial.

Neste artigo descrevemos um sistema que responde a perguntas sobre uma colecção de documentos. O sistema que apresentamos é baseado no sistema desenvolvido e avaliado no

Clef 2004 [7] que entretanto foi actualizado para ser avaliado no Clef 2005. Apresentamos e discutimos os resultados da avaliação do seu desempenho que foi feita sobre dois conjuntos de documentos: textos de dois anos de Jornais diários, o Público e a Folha de São Paulo (1994-1995); e um conjunto de documentos jurídicos que inclui as decisões do supremo Tribunal, tribunal da relação e pareceres da Procuradoria Geral da República (cerca de 100.000 documentos).

Na próxima secção a arquitectura do sistema é apresentada. Nas secções 3 e 4 são descritos em detalhe os módulos de análise sintáctica e semântica do sistema. A secção 5 apresenta o processo de representação de conhecimento utilizado. A secção 6 descreve o processo de interpretação semântico-pragmático dos documentos, com base nas análises efectuadas e na ontologia utilizada para a representação de conhecimento. Na secção 7 é demonstrado o processo de processamento e geração da resposta dada pelo sistema e na secção 8 são apresentados resultados da avaliação ao sistema. Finalmente, na secção 9 algumas conclusões e trabalho futuro são discutidos.

# 2 Arquitectura do Sistema de Pergunta-Resposta

O sistema de pergunta-resposta tem dois módulos responsáveis pelo processamento do corpus e pelo processamento das perguntas.

- Módulo de extracção de informação.

  Este módulo processa o corpus criando uma base de conhecimentos com a informação extraída dos documentos.

  O Diagrama deste módulo, ver figura 1, apresenta as várias fases do módulo de extracção:

  - Análise sintáctica: as frases dos textos da colecção de documentos são processa-

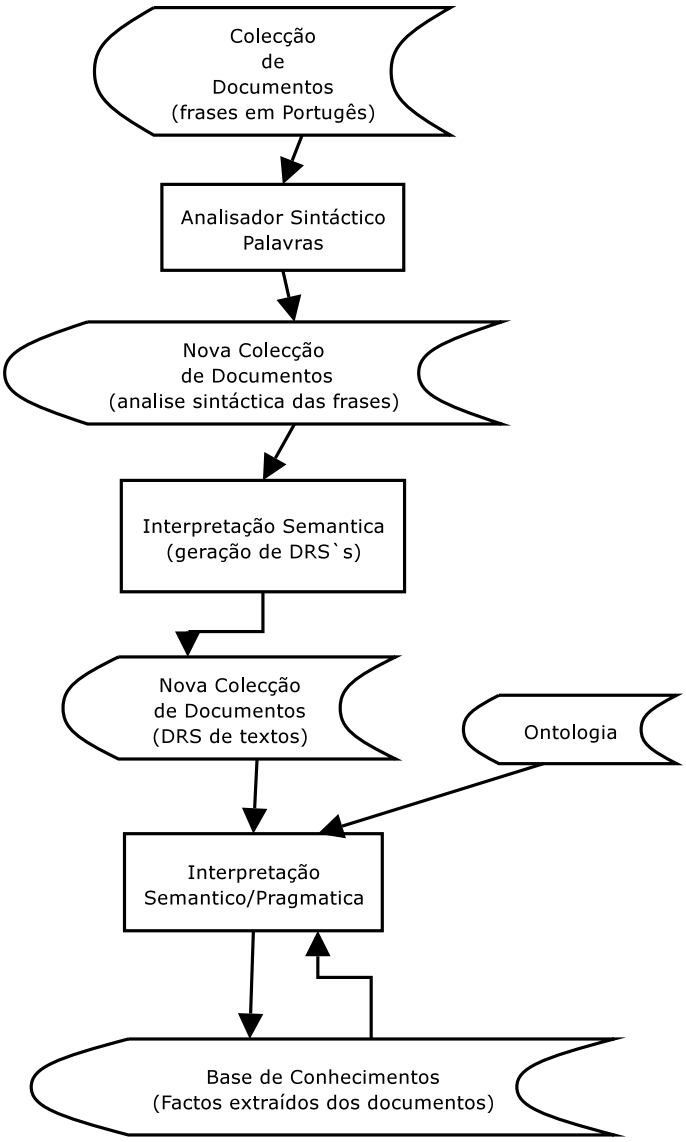


Figure 1: Processamento dos documentos da colecção

dos com analisador sintáctico Palavras[3]. O processamento é feito documento a documento, criando-se um novo documento onde cada frase é substituída pela representação da sua análise sintáctica.

Após esta fase têm-se uma nova colecção de documentos que tem o mesmo número de documentos da colecção inicial.

- Análise semântica: a nova colecção de textos, onde cada frase foi substituída pela sua análise sintáctica, é reescrita[5] originado uma nova colecção de documentos onde cada documento tem uma DRS (estrutura para representação do discurso), uma lista de referentes do discurso e um conjunto de condições. Para se poder cumprir com o requisito da avaliação do CLEF agrupam-se os referentes e condições de cada frase para que se possa saber qual foi a frase que originou o conhecimento extraído.
- Interpretação semântica e pragmática: nesta fase processa-se a colecção de documentos que se construiu na fase anterior, e tendo em conta a ontologia e o conhecimento já extraído vai-se populando uma base de dados. Esta base de dados contém instâncias da ontologia e tabelas que representam predicados de aridade variável.

- Módulo de recuperação de informação.

  Este módulo processa a pergunta em Português e retorna a resposta: uma conjunto de palavras e a identificação do documento e da frase onde encontrou a resposta.

  Na figura 2 está o diagrama deste módulo. Este módulo tem as seguintes fases:

  - Análise Sintáctica: com o analisador Palavras[3] é feita a análise sintáctica da pergunta.

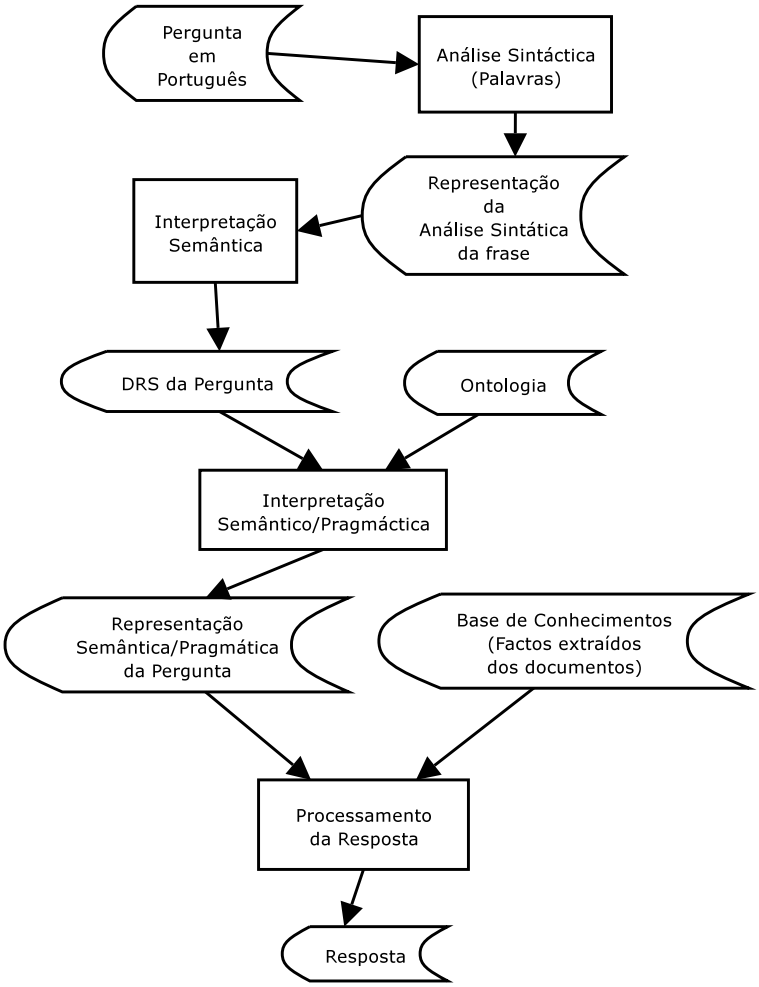


Figure 2: Processamento da Pergunta

– Análise Semântica: a partir da análise sintáctica constrói-se a estrutura do discurso que representa a pergunta, uma DRS[5] com os referentes da interrogativa marcados.

– Interpretação Semântico/Pragmática: nesta fase a representação semântica da frase, algumas condições são reescritas, tendo em conta a(s) ontologia(s), originando numa nova DRS .

– Processamento da Resposta: a representação final da pergunta é interpretada na Base de Conhecimentos com os factos extraídos da colecção de documentos

alvo. Esta interpretação é feita através da unificação das entidades do discurso da pergunta com entidades do discurso dos documentos (ver secção 7).

Como podem existir várias frases de vários documentos que suportam uma resposta à pergunta, este processo de cálculo da resposta actualmente calcula todas as respostas para a pergunta e tem um algoritmo para escolher a "melhor" resposta.

Nas próximas secções vamos apresentar com algum detalhe os processos das diferentes fases dos módulos do sistema de Pergunta-Resposta.

# 3 Análise Sintáctica

A estrutura sintáctica das frases, quer dos documentos quer das perguntas, é obtida usando o analisador analisador sintáctico de Eckhard Bick[3], devolvido no contexto do projecto VISL[1] no *Insttituto da Linguagem e Comunicação* na *Universidade de Southern Denmark*. Este analisador fornece ampla informação morfo-sintáctica sobre os constituintes e o léxico das frases. Por exemplo, quando este analisador identifica o verbo principal de uma frase, retorna, além dessa identificação, o lema do verbo (o verbo na forma infinitiva), que irá ser usado como nome de predicado pela análise semântica, evitando assim a proliferação de nomes de predicados equivalentes. Esta característica do analisador sintáctico é essencial para uma Língua como o Português que tem uma morfologia muito rica e variada.

Contudo, apesar de este analisador sintáctico não identificar as posições para argumentos vazios permite que estas sejam inferidas na interpretação semântica pois as posições de sujeito e os complementos são identificadas através de etiquetas, por exemplo: *SUBJ* e *ACC* para sujeito e complemento directo. Estas etiquetas permitem-nos, na interpretação semântica, inferir uma ordem adequada para os argumentos nos predicados.

[1] Visual Interactive Syntax Learning

Mais problemático é a resolução da ligação das preposições (pp-attachment): como decidimos ficar com a primeira análise sintáctica da frase (normalmente há mais do que uma) as alternativas à ligação proposta pela análise sintáctica são consideradas na interpretação semântico/pragmática.

Considere a frase (3.1):

> Um patologista defendeu que Jimi Hendrix morreu de asfixia após ter ingerido álcool e uma dose excessiva de barbitúricos. (3.1)

A estrutura sintáctica que resulta da análise desta frase feita pelo Palavras é a seguinte:

```
STA:fcl
=SUBJ:np
==>N:art( um  M S <arti>) Um
==H:n( patologista  M S <Hprof>)
  patologista
=P:v-fin( defender  PS 3S IND)
  defendeu
=ACC:fcl
==SUB:conj-s( que ) que
==SUBJ:prop( Jimi_Hendrix  M/F S)
  Jimi_Hendrix
==P:v-fin( morrer  PS 3S IND) morreu
==PIV:pp
===H:prp( de ) de
===P<:np
====H:n( asfixia  F S <sick>) asfixia
====N<:pp
```

```
=====H:prp( após ) após
=====P<:icl
======P:vp
=======AUX:v-inf( ter ) ter
=======MV:v-pcp( ingerir ) ingerido
======ACC:n( álcool  M S <cm-liq>)
   álcool
======CO:conj-c( e ) e
======ACC:np
=======>N:art( um  F S <arti>) uma
=======H:n( dose  F S) dose
=======N<:adj( excessivo  F S)
   excessiva
=======N<:pp
========H:prp( de ) de
========P<:n( barbitúrico  M P)
   barbitúricos
```

Esta estrutura é transformada numa representação em Prolog equivalente que se representa em baixo. É a representação em Prolog é que é guardada e é usada como entrada para a interpretação semântica.

```
sta(fcl,
    subj(np,
        n(art('um',
              'M', 'S', <arti> ), 'Um'),
        h(n('patologista',
```

```
         'M', 'S', <Hprof> ),
      'patologista')),
p(v_fin('defender', 'PS', '3S', 'IND'),
   'defendeu'),
acc(fcl,
    sub(
        conj_s('que'), 'que')),
subj(prop('Jimi_Hendrix', 'M/F', 'S'),
     'Jimi_Hendrix'),
p(v_fin('morrer', 'PS', '3S', 'IND'),
  'morreu'),
piv(pp,
    h(prp('de'),'de'),
    p(np,
      h(n('asfixia','F','S', <sick> ), 'asfixia'),
      n(pp,
        h(prp('após'), 'após'),
        p(icl,
          p(vp,
            aux(v_inf('ter'),'ter'),
            mv(v_pcp('ingerir'), 'ingerido')),
          acc(n('álcool',
                'M', 'S', <cm-liq> ), álcool'),
          co(conj_c('e'), 'e'),
          acc(np,
              n(art('um',
```

```
      'F','S',<arti>), 'uma'),
h(n('dose','F','S'), 'dose'),
n(adj('excessivo',
      'F','S'), 'excessiva'),
n(pp,
  h(prp('de'),'de'),
  p(n('barbitúrico',
       'M','P'), 'barbitúricos' , '.' )
) ) ) ) ) ) ).
```

# 4 Análise Semântica

A análise semântica, ou interpretação semântica, reescreve uma estrutura sintáctica numa estrutura de representação do discurso[5], DRS. Na realidade no sistema só representamos frases factuais, i.e, frases que envolvem só o a quantificação existencial sobre as entidades do discurso. Assim para nós uma estrutura de representação do discurso é um conjunto de referentes, variáveis existencialmente quantificadas, e um conjunto de condições, predicados de primeira ordem.

Apesar de a Teoria de Representação do Discurso[5] prever a representação de frases complexas como condicionais, e outras modalidades; ou a representação de entidades universalmente quantificadas; e prever também o tratamento de substantivos no plural; o sistema de Pergunta-Resposta, nesta fase, não representa esta tipo de frases.

Cada árvore sintáctica, representada em Prolog, é reescrita de acordo com um conjunto de regras, e integrada numa DRS.

Para possibilitar interpretações alternativas, a ligação dos sintagmas proposicionais é

feita introduzindo a condição *rel* com 3 argumentos, a preposição e duas entidades do discurso. Este predicado *rel* permite que na interpretação semântico pragmática seja feita a ligação adequada entre as duas entidades do discurso. Por exemplo, a frase 'O dono do cão', é representada pela seguinte DRS:

```
drs(
    entidades:[ A:(definido, masc, sing),
                B:(definido, masc, sing)],
    condições:[dono(A),
               cão(B),
               rel(de,A,B)]
)
```

Como se verá na próxima secção esta representação permite que na interpretação pragmática as condições sejam reescritas originando a seguinte DRS:

```
drs(
    entidades:[ A:(definido, masc, sing),
                B:(definido, masc, sing)],
    condições:[pertence(A,B),
               pessoa(A),
               cão(B)]
)
```

Para exemplificar o resultado da reescrita de árvores sintácticas em DRS's, em baixo apresenta-se a DRS que resulta da reescrita da frase (3.1):

```
drs (entidades:[ A: (indefinido, masc, sing),
                 B: (definido, masc/fem, sing),
```

```
                C: (definido, fem, sing),
                D: (definido, masc, sing),
                E: (indefinido, fem, sing) ],
    condições:[ patologista (A),
                defende(A,B),
                nome(B, 'Jimmy Hendrix'),
                morrer(B),
                rel (de, B, C),
                asfixia(C),
                rel (após, C, D),
                ingestão(D),
                álcool(D),
                dose(D),
                excessivo(D),
                rel(de, D, E),
                barbitúricos(E)])
```

As perguntas feitas pelo utilizador também são interpretadas e reescritas numa DRS, neste caso as interrogativas aparecem marcadas nas entidades do discurso. Por exemplo, considere a seguinte questão:

Como morreu Jimi Hendrix? (4.1)

Esta frase é representada pela seguinte estrutura do discurso:

```
drs(
    entidades:[F:(definido, masc/fem, sing),
               G: interrog(que), masc, sing]
```

```
    condições:[morrer(F),
               nome(F,'Jimmy Hendrix'),
               rel(de,F,G)]
)
```

Esta representação é obtida porque "Como" é interpretado como "de que". Esta opção resulta numa sobre-geração de regras, obtendo-se assim várias alternativas para a relação *rel*. Na interpretação semântico/pragmática e no processamento da resposta, esta representação da frase (4.1) pode ser unificada com a representação da frase (3.1), ou poderia ser unificada com a representação de uma frase como: "Jimmy Hendrix morreu afogado".

# 5 Ontologia e Representação dos factos extraídos

Para representar a ontologia e os factos extraídos dos documentos, usa-se uma extensão à programação em Lógica, o ISCO[1, 2], que permite um acesso transparente, no Prolog, a bases de dados. Esta tecnologia é fundamental para viabilizar o sistema desenvolvido, pois a dimensão da base de dados é um problema real, temos mais de 9 milhões de referentes do discurso após o processamento de dois anos do jornal "O Público".

As bases de dados são definidas em ISCO a partir de uma ontologia.

O sistema usa 3 tipos de ontologia:

- uma ontologia construída pela equipa com o objectivo de modelar algum tipo de conhecimento geral como Geografia Humana e Datas;

  Este tipo de conhecimento é importante para extrair factos de forma a que se possam responder a perguntas sobre lugares. Esta ontologia define locais (cidades, países, continentes, conselhos, distritos, aldeias, etc) e relações entre locais (é capital do país, pertence ao conselho, pertence ao distrito, etc).

- uma ontologia especifica para o domínio Jurídico;

  Esta ontologia tem entidades e propriedades definidas para o domínio jurídico e é utilizada com o objectivo de procurar analisar o desempenho do sistema a responder a perguntas sobre este domínio.

  Na realidade, quando os documentos não são específicos de um domínio o uso desta ontologia não melhorou o desempenho do sistema.

- uma ontologia gerada automaticamente a partir do corpus (colecção de documentos)[9, 8];

  Esta ontologia, apesar de ser muito simples e com pouca estrutura – tem muito poucas relações *é_um* (is_a) – tem que existir para consigamos extrair factos dos textos.

A ontologia pode ser definida directamente em ISCO, ou em OWL (Ontology Web Language) e depois transformada em ISCO [8].

Actualmente o que é guardado pelo módulo de extracção de conhecimento são factos que são guardados como termos de tabelas de uma Base de Dados.

Por exemplo a frase (3.1) que tem a representação semântica apresentada na página 14, se após a interpretação semântica/pragmática mantivesse a mesma representação, daria origem à actualização dos tuplos constituídos pelos argumentos nas tabelas com o nome do predicado.

Antes de actualizar a base de conhecimentos as expressões de lógica de primeira ordem, DRS's, são *skolamizadas*, i.e cada variável existencialmente quantificada é substituída por uma constante diferente.

Assim, para a frase (3.1), os seguinte tuplos, entre outros, seriam adicionados à base de dados:

- `(123,''Jimmy Hendrix'')` acrescentado à tabela *nome*

Mesmo que já existam outros identificadores (entidades do discurso) com este nome, por exemplo, se o tuplo `(23,''Jimmy Hendrix'')` já pertencesse à tabela *nome*, o sistema não tentaria resolver a referência. Este é um aspecto que pensamos tratar no futuro.

- `(123)` acrescentado à tabela *morrer*
- `(124)` acrescentado à tabela *asfixia*
- `rel(de,123,124)` acrescentado à tabela *rel*

No processamento dos documentos, ao contrário do processamento das respostas, o sistema compromete-se com a primeira interpretação de cada frase. Esta opção, feita tendo em conta a complexidade temporal e espacial do processo de aquisição, compromete, nalguns casos, a possibilidade de obter a interpretação correcta. No entanto, o facto de falharmos a interpretação correcta para algumas frases não parece comprometer o desempenho do sistema, a sua capacidade de responder, pois a informação contida numa colecção de jornais é muito redundante.

Para que o sistema possa restringir as suas respostas à informação veiculada por uma só frase dum documento, para cada entidade do discurso guarda-se a informação sobre o documento e a frase do documento numa tabela da base de dados. Por exemplo, o tuplo `(123,'publico/publico95/950605/005',4)` seria acrescentado à tabela *referido_em*.

# 6 Interpretação Semântico/Pragmática

A interpretação Semântico/Pragmática deve reinterpretar a representação semântica da frase à luz da ontologia considerada, de forma a obter o conjunto de factos que representam a informação veiculada pela frase (ou pergunta).

O processo que faz a interpretação semântico/pragmática recebe uma estrutura de representação do discurso, uma DRS, e vai interpretá-la numa base de conhecimentos com as regras obtidas a partir da ontologia, e a informação contida na base de dados.

Para conseguir uma boa interpretação para a frase, a nossa estratégia é procurar a melhor explicação para que a forma lógica da frase seja verdadeira na base de conhecimentos para a interpretação semântico/pragmática. Esta estratégia para a interpretação pragmática foi inicialmente proposta em [4].

A Base de Conhecimentos para a interpretação pragmática é construída a partir da descrição da Ontologia em ISCO. A inferência na Base de Conhecimentos usa abdução e variáveis restringidas a domínios finitos (GNU Prolog Finite Domain (FD) constraint solver).

Considere a seguinte frase:

"A. conduzia com uma taxa de alcoolemia de 2.15."

que pela análise semântica é transformada na seguinte estrutura, uma DRS com 4 entidades do discurso e um conjunto de condições:

```
drs(
    entidades:[A:(definida,masc,sing),
               B:(indefinido,fem,sing),
               C:(indefinido,fem,sing),
               D:(definida,masc,sing)]
    condições:[nome(A, 'A.'),
               conduzir(A),
               rel(com,A,B),
               taxa(B),
               rel(de,B,C),
```

```
                alcoolemia(C),
                rel(de,C,D),
                número(D,2.15)]
)
```

O processo de interpretação semântico/pragmática usando a informação da ontologia vai reescrever esta DRS na seguinte:

```
drs(
    entidades:[BA:(definido,masc,sing),
               :(definido,masc,sing)]
    condições:[nome(A, 'A.'),
               pessoa(A),
               conduzir(A,_,_,_,B),
               taxa\_alcoolemia(B,2,15)]
)
```

A interpretação de *rel(com,A,B)* como *conduzir(A,_,_,_,B)* é possível porque a nossa ontologia tem uma propriedade *conduzir* que relaciona pessoas que conduzem num intervalo de tempo com uma taxa de alcoolemia no sangue.

Neste processo a tarefa de criar identificadores únicos para (skolamizar) as variáveis existencialmente quantificadas da DRS final também é fundamental para que na base de dados se garanta a consistência da informação extraída.

Este passo do processo de interpretação semântico/pragmática é uma das fontes de problemas do sistema de Pergunta-Resposta pois desta forma não vamos impedir que um individuo tenha vários identificadores diferentes. O problema de unificar entidades do discurso com identificadores já existentes na base de dados pode originar que indivíduos diferentes tenham o mesmo identificador, o que seria uma situação indesejável.

Actualmente uma Base de Dados contém toda a informação extraída de uma colecção de documentos. No entanto usando o ISCO é possível usar as diferentes bases de dados, correspondentes a colecções de documentos diferentes, durante o processo de interpretação pragmática e durante o processamento das respostas.

Na extracção da informação, o processo de aquisição de dados, ainda não tem em atenção o facto de a representação semântica poder ser ambígua, i.e. existir mais do que uma interpretação para a frase. De momento o nosso sistema na aquisição de informação usa só a primeira interpretação. No entanto no processamento da resposta o sistema já tem em conta as diferentes interpretações possíveis para a pergunta no cálculo da resposta.

Os nossos esforços na avaliação do sistema mostram que o facto de se perderem algumas representações adequadas para algumas frases pode não ter consequências no desempenho do sistema, pois a informação contida em bases de documentos muito grandes é redundante e mesmo que se perca o conteúdo de algumas frases é sempre possível obter a informação de outras.

# 7 Processamento da Resposta

O processamento da resposta é feito em dois passos:

1. Identificação do referente da base de dados que unifica com o referente afectado pelo pronome interrogativo na pergunta.

2. Recolha das propriedades do referente identificado no passo anterior de forma a construir a resposta.

Para exemplificarmos o processamento da resposta, considere a seguinte pergunta:

"Quem cometeu um homicídio por negligência por conduzir alcoolizado?

Esta pergunta é representada pela seguinte DRS depois de se ter feito a análise sintáctica e análise semântica.

```
drs(
    entidades:[A:(quem,masc/fem,sing),
               B:(indefinido,masc,sing),
               C:(indefinido,fem,sing),
               D:(indefinido,masc,sing)],
    condições:[cometeu(A,B),
               homicídio(B),
               rel(por,B,C),
               negligência(C),
               rel(por,A,D),
               alcoolizado(D),
               conduzir(D)]
)
```

A interpretação semântico/pragmática desta pergunta é feita usando a ontologia de conceitos e permite obter a seguinte DRS como representação do seu significado:

```
drs(
    entidades:[A:(quem,masc/fem,sing),
               B:(indefinido,masc/fem,sing/plu),
               C:(definido,fem,sing),
    condições:[homicídio(A,B),
               pessoa(A),
               pessoa(B),
               conduzir(A,_,_,_,C),
```

```
                taxa_alcoolemia(C),C>0.5]
)
```

- Para executar o primeiro passo:

  Mantém-se as variáveis dos referentes, e começa-se por provar as condições da DRS na base de conhecimentos com os factos extraídos dos documentos. Se a condições forem verdadeiras na base de conhecimentos, os referentes do discurso serão unificados com identificadores de indivíduos da base de dados.

- O passo seguinte é recolher as palavras que constituem a resposta:

  Neste caso deve-se recolher todas as condições que envolvem a entidade do discurso *A* e escolher a que melhor caracteriza a entidade. Indiscutivelmente a primeira escolha deve recair sobre o a existência de uma condição com o predicado *nome* (nome(A,Nome)).

  Mas nem sempre é tão simples encontrar a resposta adequada à pergunta. Sobretudo, quando a pergunta é do tipo:

  - Que crimes cometeu o X?
  - Quantos habitantes tem Kalininegrado?
  - Qual a nacionalidade da Miss Universo?
  - Quem é Flavio Briatore?

  Para escolher a melhor resposta a dar a uma pergunta de entre as possíveis o nosso sistema têm um algoritmo que leva em conta as categorias sintácticas das palavras que podem ser utilizadas nas respostas e tem em conta que não deve repetir palavras que estão na pergunta.

As perguntas sobre lugares e datas têm um tratamento especial que envolve a consulta de dados sobre lugares ou datas.

No processamento da resposta também temos que ter em conta a o número de respostas que pretendemos calcular. Para uma pergunta podem existir várias respostas diferentes dependendo da informação contida na base de conhecimentos. No CLEF 2005 optamos por calcular todas as respostas possíveis para cada pergunta e escolher a mais frequente.

Uma imposição do CLEF é que a resposta seja suportada por uma só frase de um documento, esta restrição é satisfeita pelo nosso sistema impondo, no passo 1, que todas as entidades do discurso da drs da pergunta tenham sido introduzidas pelo mesmo documento na mesma frase. O predicado *referido_em* permite obter essa informação (ver página 18).

# 8 Avaliação

A avaliação do sistema de pergunta-resposta foi efectuada no âmbito do CLEF – Cross Language Evaluation Forum – de 2004 e 2005. Neste fórum, um conjunto de perguntas elaborada por um júri é fornecida ao sistema, sendo as respostas avaliadas pelo mesmo júri.

O nosso sistema fornece como *output* uma expressão em Língua Natural que deverá conter a resposta à pergunta e a referência ao documento/frase que suporta essa resposta.

Os testes foram realizados a partir de 200 perguntas e a as respostas foram avaliadas de acordo com a seguinte classificação:

- resposta correcta e bem suportada.

  Indica que a resposta estava correcta e suportada pelo documento adequado.

  25% das respostas tiveram esta classificação.

- resposta correcta mas não suportada.

Indica que a resposta estava correcta mas não estava suportada pelo documento indicado.

2% das respostas tiveram esta classificação.

- resposta incorrecta mas o documento/frase continha a resposta adequada.

  Indica que a resposta não estava correcta mas o sistema foi capaz de seleccionar o documento adequado.

  18% das respostas tiveram esta classificação.

- resposta incorrecta e a frase também não continha a informação devida.

  9% das respostas tiveram esta classificação.

- o sistema não respondeu à questão.

  46% das perguntas não obtiveram resposta.

Estes resultados apontam para cerca de 25% de respostas correctas.

A análise dos 46% de casos em que o sstema não conseguiu obter uma resposta demonstrou que a principal causa dos erros era devida a falta de conhecimento do sistema: análises sintácticas erradas (tanto das perguntas, como das frases do documento), falta de informação sobre sinónimos e, acima de tudo, o uso de uma ontologia incompleta. De facto, a maioria dos problemas está relacionado com a impossibilidade de efectuar a correcta interpretação semântico-pragmática, devido a lacunas na ontologia que está a ser utilizada. O processo de inferência utilizado é baseado na ontologia de representação de conhecimento. Por exemplo, é fundamental saber o que são locais, pessoas, datas, etc. – se não houver informação que indique que "Lisboa" é uma cidade, então não será possível responder a perguntas sobre a "cidade de Lisboa" ou a "capital de Portugal". Neste sentido é fundamental continuar a desenvolver metodologias que permitam a construção e a integração automática de ontologias.

No entanto, e tendo em conta os resultados obtidos pelos diversos sistemas de pergunta-resposta concorrentes ao CLEF, pode-se afirmar que o sistema proposto se encontra entre os que obtiveram melhores resultados e possui um potencial para melhorias muito interessante.

Em concreto, o sistema é bastante fiável nas respostas que apresenta, tendo uma precisão bastante alta: das 108 respostas apresentadas, em somente 18 o sistema seleccionou informação incorrecta – cerca de 84% de precisão. Mesmo nas respostas incorrectas, o utilizador poderá confirmar a informação fornecida pelo sistema, pela simples leitura da frase de suporte à resposta.

# 9 Conclusões e Trabalho Futuro

Este artido descreve uma proposta de arquitectura para um sistema de pergunta-resposta para a Língua Portuguesa.

O sistema proposto utiliza técnicas de processamento de Língua Natural para criar uma base de conhecimento com a informação veículada pelos diversos documentos. As interrogações dos utilizadores são analisadas pelas mesmas ferramentas e são efectuadas inferências lógicas sobre a base de conhecimentos, tentando encontrar a resposta adequada. O processo de inferência é efectuado recorrendo a um ambiente de programação em lógica e ao motor de inferência da linguagem Prolog.

Os principais problemas do sistema estão relacionados com erros nos diversos módulos de PLN – análise sintáctica e análise semântica – e com a falta de cobertura da ontologia utilizada. De facto, a necessidade de uma ontologia adequada ao domínio dos documentos a processar é o maior problema do sistema proposto.

Como trabalho futuro, pretende-se propôr novas metodologias para tentar ultrapassar parcialmente estes problemas. Uma das linhas de investigação estará relacionada com a construção automática de ontologias a partir de documentos em Língua Natural; a in-

tegração automática de diversas ontologias será outra das áreas a investigar. O aperfeiçoamento dos módulos de PLN existentes (sintaxe, semântica) é outra das áreas a necessitar de bastante trabalho e pretendemos, ainda, explorar o problemas das referências anafóricas nos documentos, de modo a diminuir o número de referentes distintos na base de conhecimentos e a permitir inferências mais complexas.

# Referências


[1] Salvador Abreu. Isco: A practical language for heterogeneous information system construction. In *Proceedings of INAP'01*, Tokyo, Japan, October 2001. INAP.

[2] Salvador Abreu, Paulo Quaresma, Luis Quintano, and Irene Rodrigues. A dialogue manager for accessing databases. In *13th European-Japanese Conference on Information Modelling and Knowledge Bases*, pages 213–224, Kitakyushu, Japan, June 2003. Kyushu Institute of Technology. To be published by IOS Press.

[3] Eckhard Bick. *The Parsing System "Palavras". Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus University Press, 2000.

[4] Jerry Hobbs, Mark Stickel, Douglas Appelt, and Paul Martin. Interpretation as abduction. Technical Report SRI Technical Note 499, 333 Ravenswood Ave., Menlo Park, CA 94025, 1990.

[5] Hans Kamp and Uwe Reyle. *From Discourse to Logic:An Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Dordrecht: D. Reidel, 1993.

[6] M-F Moens. Interrogating legal documents: The future of legal information systems? In *Proceedings of the JURIX 2003 Workshop on Question Answering for Interrogating Legal Documents*, pages 19–30, Utrecht, Netherlands, 2003. Utrecht University.

[7] Paulo Quaresma and Irene Rodrigues. Using dialogues to access semantic knowledge in a web legal IR system. In Marie-Francine Moens, editor, *Procs. of the Workshop on Question Answering for Interrogating Legal Documents of JURIX'03 – The 16th Annual Conference on Legal Knowledge and Information Systems*, Utrecht, Netherlands, December 2003. Utrecht University.

[8] José Saias. Uma metodologia para a construção automática de ontologias e a sua aplicação em sistemas de recuperação de informação – a methodology for the automatic creation of ontologies and its application in information retrieval systems. Master's thesis, University of Évora, Portugal, 2003. In Portuguese.

[9] José Saias and Paulo Quaresma. Using nlp techniques to create legal ontologies in a logic programming based web information retrieval system. In *Workshop on Legal Ontologies and Web based legal information management of the 9th International Conference on Artificial Intelligence and Law*, Edinburgh, Scotland, June 2003.