Um sistema de pesquisa de informação para bases de texto em Português

Paulo Quaresma and Irene Pimenta Rodrigues
and Gabriel Lopes
email:(pq,ipr,gpl)@di.fct.unl.pt
AI Center, Departamento de Informática
Universidade Nova de Lisboa
Quinta da Torre, 2825 Monte da Caparica,
Portugal

Teresa Almeida and Elsa Garcia and Ana Lima
email: (talmeida,egarcia,analima)@pgr.pt
Procuradoria Geral da República
Rua da Escola Politécnica, 140
1294 Lisboa Codex,
Portugal

Resumo

Neste artigo descrevemos um sistema de pesquisa de informação com interface para a web para bases de textos em Português cosntituída pelos pareceres da Procuradoria Geral da República.
O engenho de pesquisa é especializado para o Português, tendo em conta informação lexical, sintáctica e alguma semântica. A informação semântica é obtida através de um thesaurus com expressões jurídicas relacionadas pelas relações: equivalente, relacionado com, específica, generaliza. Esta relações entre os termos jurídicos são utilizadas para manipular as questões do utilizador e as respostas do sistema. Tenta-se expandir as questões de forma a ter em conta os termos relacionados e os termos específicos, e tenta-se colapsar as respostas usando os termos genéricos de forma a obter respostas mais gerais.
O sistema tenta manter o contexto de interação, sempre que possível, juntando a nova questão do utilizador às anteriores, permitindo o refinamento das questões de uma forma muito amigável.
O sistema também permite a representação de conhecimento jurídico usando uma linguagem de programação em lógica tornando possivel a verificação das regras usando os documentos específicos da base de textos.

1. Introdução

Neste artigo apresentamos um sistema inteligente de pesquisa de informação que tem uma interface WEB para uma base de textos com textos jurídicos, os pareceres da Procuradoria Geral da República. O objectivo principal desta inteface é permitir a pesquisa de documentos da base de textos. O utilizador coloca uma questão e deve obter como resposta o conjunto de documentos da base de textos que satisfazem a questão, bem como algumas sujestões para refinamento posterior.
A base de textos tem um conjunto de documentos (7000) que estão estruturados em secções que incluem: conclusões, texto integral, informação administrativa e descritores (classificação com termos jurídicos).
Este conjunto de documentos foi processado com o SINO, um engenho de pesquisa para bases de textos jurídicos (Greenleaf, Mowbray and King 1997). Alterou-se o SINO de forma a poder lidar com algumas articularidades do Portugês, stop words, plurais, verbos, e sinónimos.
De forma a poder lidar com estas particularidades, usa-se um dicionário lexical que foi construído pela nossa equipa na FCT/UNL, aumentado com o vocabulário dos pareceres da PGR. De forma a poder dar um tratamento correcto aos substantivos e aos verbos utiliza-se um etiquetador automático para o Português, que marca a categoria lexical das

palavras no seu contexto, substantivo ou verbo. O etiquetador e o dicionário são utilizados no processo de indexação dos textos com o SINO e no processamento as questões do utilizador. Desta forma é possível evitar a referência a documentos devido a ambiguidade lexical (substantivos que se escrevem como algumas formas verbais).
O nosso sistema é hibrido, no sentido em que usa diferentes fontes de conhecimento para construir as respostas ao utilizador e as propostas de refinamento, em particular usa o thesaurus de termos jurídicos e a base de conhecimentos com as regras jurídicas. Os pareceres da PGR têm uma secção com a classificação jurídica, que contém um conjunto de termos jurídicos pré-definidos, descritores, que resultam da análise jurídica feita pela PGR. Os nossos parceiros da PGR estudaram os termos jurídicos, descritores, de forma a construir um thesaurus com as relações: *equivalente*, *específica*, *generaliza* e *relacionado com*.
Dada a classificação dos documentos, o thesaurus e um conjunto de documentos, é possivel agrupar conjuntos de documentos que têm o mesmo descritor ou um descritor relacionado (pela relação equivalente, generaliza ou relacionado com). Esta é uma forma de propôr possíveis refinamentos das questões do utilizador. Esta forma de refinamento é particularmente interessante para os utilizadores que não conhecem o conjunto dos descritores utilizados pela PGR na classificação dos documentos, estudantes de direito e e público em geral.
O thesaurus de termos jurídicos é também utilizado na expansão das questões do utilizador. Dada uma palavra ou uma expressão que exista no thesaurus, a palavra ou expressão é expandida num conjunto de expressões, usando as relações do thesaurus.
Para obter uma visão estruturada de todos os documentos na base de textos também se utiliza o thesaurus, podendo-se visualizar a estrutura dos descritores e todos os documentos que referem um determinado descritor. Também é possivel obter uma visão estruturada dos documentos utilizando a informação dos documentos que referem outros na base de textos.
Estas vistas da base de textos são muito úteis para os utilizadores sem conhecimento jurídico, público em geral.

Por enquanto,a nossa interface para o web só suporta questões em expressões Booleanas tipo SQL, permitindo ao utilizador procurar a ocorrência de uma palavra, ou de uma expressão, no documento todo ou só nalguma secção. O sistema tenta manter o contexto da interrogação nos dois tipos de questões e trata a sequência pergunta/resposta como um diálogo entre dois agentes inteligentes.

2. O Thesaurus com os termos Jurídicos

Todos os pareceres da PGR são sujeitos a uma análise jurídica da qual resulta a atribuição de um ou mais descritores para classificar cada documento. Os descritores são escolhidos do conjunto de descritores que tem sido aumentado ao longo dos anos, por ora existem mais de 6000 descritores. Quando este projecto se iniciou foram desenvolvidos esforços no sentido de agrupar e relacionar todos os descritores. Neste momento, esta ainda é uma das principais tarefas do projecto, pois o conhecimento jurídico contido nestas espressões (descritores) pode ser explorado de diferentes formas.

Nesta secção apresenta-se primeiro a organização do thesaurus e depois apresenta-se algumas das formas como é utilizado pelo sistema.

As relações no thesaurus

Para cada descritor tentou-se listar os descritores que estão numa das seguintes relações:

- É equivalente a
  Ex: **lei** é equivalente a **norma**
- É generalizado por
  Ex: **primeiro ministro** é generalizado por **ministro**
- É especificado por
  Ex: **acidente** é especificado por **acidente de viação**
  desastre
- É relacionado com
  Ex: **deserção** é relacionado com **acidente de viação**
  exército

As propriedades destas relações são:

- ***Equivalente***
  1. A **é equivalente a** B => B **é equivalente a** A
  2. A **é equivalente a** B**,** B **é equivalente a** C **=>** A **é equivalente a** C
- ***Generalizado***
  1. A **é especificado por** B => B **é generalizado por** A
  2. A **é generalizado por** B, B **é equivalente a** C => A **é generalizado por** C
  3. * A **é generalizado por** B, B **é relacionado com** C => A **é generalizado por** C
- *Especificado*
  1. A **é generalizado por** B => B **é especificado por** A
  2. A **é especificado por** B, B **é equivalente a** C => A **é especificado por** C
  3. * A **é especificado por** B, B **é relacionado com** C => A **é especificado por** C
- *Relacionado*
  1. A **é relacionado com** B => B **é relacionado com** A

Dada uma lista inicial de descritores, as suas relações e as propriedades das relações, é possível calcular o fecho obtendo uma nova lista de descritores que pode ser verificada por especialistas da PGR para analisarem as consequências das definições iniciais. Este tem sido o método utilizado nesta tarefa de construção do thesaurus.

Uso do thesaurus

O conhecimento jurídico contido no thesaurus está a ser utilizado em 3 módulos diferentes:

- Processamento das perguntas
- Cálculo das propostas de refinamento
- Visionamento dos documentos

Estes 3 módulos vão ser descritos abaixo.

Processamento das Perguntas

O conhecimento contido no thesaurus é usado para expandir as perguntas do utilizador. Sempre que uma pergunta especifica um descritor, a pergunta é expandida com todos os descritores que são equivalentes ou mais específicos ou relacionados com o descritor inicial.

Suponhamos que um utilizador quer ser informado sobre os pareceres sobre acidentes. O utilizador pode formular a sua pergunta de diferentes formas, por exemplo:

- Acidente
- Texto_Integral(acidente) or Conclusão(acidente)
- Descritor(acidente)

O sistema expande o descritor em cada pergunta usando o thesaurus adicionando os seguintes descritores

- todos os descritores equivalentes a acidente
- todos os descritores que estão no fecho transitivo da relação:
  acidente é especificado por
- todos os descritores relacionados com acidente

Neste caso, a palavra acidente será substituida por: "acidente OR acidente de viação OR deserção OR desastre"
Como resultado deste procedimento os utilizadores podem obter documentos onde a palavra *acidente* não está presente em nenhuma secção, mas o documento será potencialmente sobre acidentes. Os casos onde se obtêm piores resultados (selecção de documentos que não são sobre acidentes) são aqueles que resultam da expansão com a relação "relacionado com". Mesmo com a adição de documentos irrelevantes os nossos utilizadores ainda acham este procedimento muito útil pois quando o sistema oferece propostas de refinamento é fácil eliminar os documentos irrelevantes.

Cálculo das propostas de refinamento

Sempre que é selecionado um conjunto muito grande de documentos, calculam-se propostas de refinamento. Estas propostas são apresentadas como uma lista de descritor-número de documentos. O utilizador pode selecionar um ou mais itens da lista e lançar uma nova pergunta cujo significado é a pergunta inicial e a conjunção dos descritores seleccionados.
Uma vez que a lista de propostas de refinamento pode ser muito grande, pode ser necessário colapsar conjuntos de descritor-número de documentos. O procedimento para colapsar uma lista de descritores-número de documentos é o seguinte:

Dado D1, o conjunto inicial de pares descritor-{conjunto de documento seleccionado}, o novo conjunto D2 é constuído:

- incluindo em D2 um par $desc_n$-S_n
  if há dois pares $desc_i$-S_i, $desc_j$-S_j em D1 tal que os descritores $desc_i$ e $desc_j$ verificam as relações: {$desc_i$ *é especificado por* $desc_n$ e $desc_j$ *é especificado por* $desc_n$} **e** $S_n=S_i+S_j$.
  Or if há dois pares $desc_n$-S_{ni}, $desc_j$-S_j em D1 tal que os descritores $desc_n$ e $desc_j$ verificam: $desc_j$ *é especificado por* $desc_n$ **e** $S_n=S_{ni}+S_j$.
  Or if o par $desc_n$-S_n está em D1 e não há nenhum descritor em D1 ou D2 que possa ser relacionado por: *é especificado por* ou *é generalizado por*.
- Substituir o par $desc_n$-S_m em D2 pelo par $desc_n$-S_n
  if há um par $desc_i$-S_i em D1 tal que os descritores $desc_i$ e $desc_n$ verificam:{$desc_i$ *é especificado por* $desc_n$} **e** $S_n=S_i+S_m$.

Este procedimento pode ser aplicado a um conjunto de pares descritor-{conjunto de documentos seleccionados} até que se obtenha um conjunto de pares com o cardinal pretendido.
Com este procedimento as propostas para refinamento das perguntas correspondem a grupos de documentos que estão juridicamente relacionados e classificados. Esta forma de fornecer as propostas de refinamento é muito útil para os nossos utilizadores pois fornece palavras que correspondem a conceitos jurídicos importantes que a maioria dos utilizadores desconhece ou não tem presente na elaboração da pergunta.

Exemplo

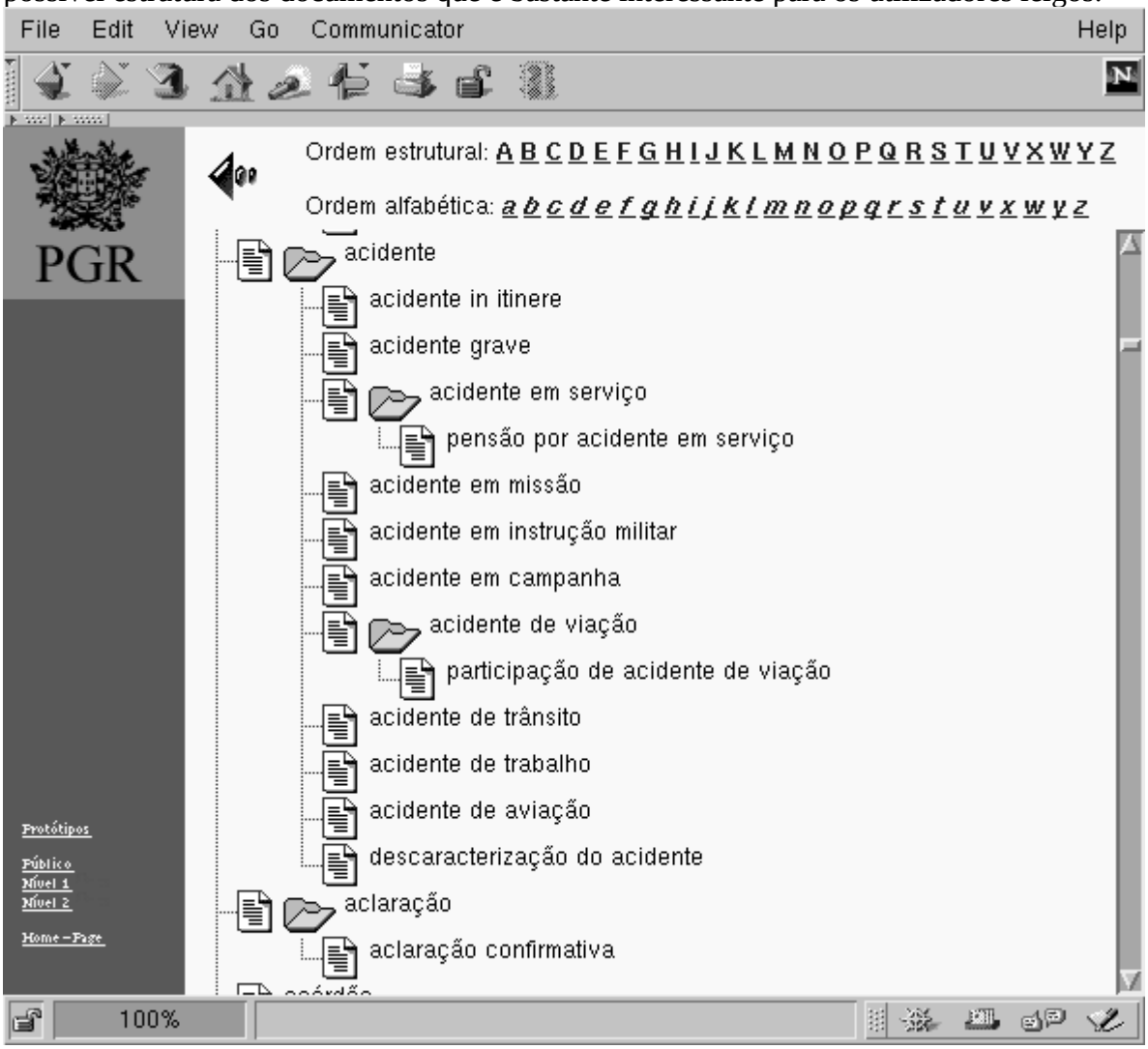
Suponhamos que o utilizador pretende ser informado sobre pareceres sobre acidentes:
- acidentes?

O sistema expande a pergunta usando o thesaurus e pesquisa com todos os descritores relacionados ou mais específicos. Por exemplo, neste caso irá pesquisar com a pergunta: "acidente OR acidente de viação OR deserção OR &..."

A seguir na resposta, o thesaurus é utilizado para colapsar o conjunto de textos seleccionados em classes relacionadas com os termos do thesaurus:
- X documentos sobre acidentes;
- Y documentos sobre deserção;
- Z documentos sobre danos civis;
- W documentos sobre danos criminais;
- etc ...

Visionamento dos Documentos

A estrutura do thesaurus permite-nos visionar o conjunto de todos os documentos agrupados pelos termos jurídicos que os classificam obtendo-se uma visão de uma possível estrutura dos documentos que é bastante interessante para os utilizadores leigos.



3. Contexto das Perguntas

Antes de processar as perguntas do utilizador o sistema tenta obter o contexto da interacção. Esta tarefa é feita seguindo o procedimento:

1. Juntar a pergunta à anterior (se existir). Esta operação é feita usando o operador AND;
2. Processar a nova pergunta (após junção com as anteriores) como foi descrito na secção anterior;
3. Se o resultado é vazio, as perguntas não são relacionadas e não devem ser "juntas". Se o resultado não é vazio, então é possível que a nova pergunta pretenda refinar a anterior ou não, e o sistema deve entendê-la assim, i.e. deve fornecer duas respostas, uma em que entende a pergunta como o refinar da anterior, e outra em que entende a pergunta como independente das anteriores.

Por exemplo suponha-se que o utilizador faz a seguinte pergunta:

- Acidente

Depoi da resposta do sistema o utilizador perguta:

- Drogas

Neste ponto o sistema tenta juntar as duas perguntas:

- Acidentes AND drogas

Como o resultado não é um conjunto vazio de documentos, os sistema dá ambas as respostas (drogas; acidentes AND drogas), permitindo ao utilizador selecionar o conjunto de documentos que deseja.

Usando esta estratégia, o sistema é capaz de de inferir e antecipar as perguntas do utilizador e actuar de uma foma amigável.

4. Modelação de Conhecimento Jurídico

Neste projecto pretende-se poder modelar conhecimento jurídico. De forma a poder tratar este problema representa-se a legislação usando programação em lógica e usamos o Prolog como engenho de inferência.

Como engenho de inferência usou-se inicialmente o YSH, construído no âmbito do projecto AustLLI (Greenleaf, Mowbray, and van Dijk, 1995), mas nós necessitamos de um engenho mais potente que possa por exemplo modelar raciocinio não monótono.

Neste momento ainda só temos um protótipo para representar legislação que define quando é que uma pessoa tem direito a uma pensão por serviços excepcionais.

Como trabalho futuro pretende-se testar se um determinado documento satisfaz uma dada especificação de uma lei. Esta tarefa é realizada através do recurso a um ambiente de programação em lógica em que se tenta provar um determinado predicado. As interrogações são transformadas em representações semânticas (por exemplo, DRS) e é verificado se cada documento suporta a referida estrutura semântica. Como, neste momento, ainda não temos a representação semântica dos documentos utilizamos a pesquisa booleana para responder às

questões.

Exemplo

Nesta secção será apresentado um exemplo acerca da legislação que define quando é que alguém tem direito a uma pensão por serviços excepcionais prestados ao País. Por motivos de espaço, este exemplo é uma versão bastante simplificada do legislação completa.

pensão(X) <- art31(X), art32(X).

art31(X) <- acção_excepcional(X), art31a(X).
art31(X) <- acção_excepcional (X), art31b(X).

art31a(X) <- acção_excepcional _local_guerra(X).
art31a(X) <- acção_abnegada_e_corajosa(X).
art31a(X) <- alto_serviço_País_ou_humanidade(X).

art31b(X) <- ferido_ou_falecido(X), acto_humanitário(X).
art31b(X) <- ferido_ou_falecido (X), acto_dedicação_causa_pública(X).

art32(X) <- demonstra_respeito_direitos_humanos(X), respeito_dignidade_País(X).

acção_excepcional(X) <- acção_beneficia_País(X,A),
acção_correcta_topologia(X,A),
acção_sem_remuneração(X,A),
acção_além_funções(X,A).

acção_correcta_topologia(X,A) <- acção_serve_interesses_nacionais(X,A),
acção_pressupõe_grande_disponibilidade(X,A).

acção_excepcional_local_guerra(X) <- acção_local_guerra(X,A),
acção_além_padrões_militares(X,A).

acção_além_padrões_militares(X,A) <-
acção_excepcional_padrões_administrativos(X,A).
acção_além_padrões_militares(X,A) <-
acção_defende_outras_vidas(X,A).

Suponhamos que se pretende testar estas regras de inferência sobre um documento específico. Par tal deve-se tentar provar o predicado: *pensão(X)*.
De modo a provar este predicado será necessário tentar provar que a pessoa realizou uma acção excepcional que, por sua vez, necessita de ser uma acção que beneficie o País (além de outras condições). Como não existe, de momento, representação semântica que permita testar estas situações, o problema é “semi-resolvido” através do recurso à pesquisa de expressões que possam descrever este tipo de acções. Por exemplo:

acção_defende_outras_vidas (X,A)<- query "salvar vida"
acção_excepcional_padrões_administrativos (X,A)<- query "além do dever" and "militar".
acção_local_guerra(X,A)<- query "guerra"

Como se pode verificar esta tarefa é bastante complexa e necessita de uma descrição do domínio bastante completa. Como trabalho futuro, pretendemos completá-la para o domínio apresentado nesta secção.

5. Bibliography

- [AKPT91] James Allen, Henry Kautz, Richard Pelavin, and Josh Tenenberg. Reasoning about Plans. Morgan Kaufman Publishers, Inc., 1991.
- [AP96] José J. Alferes and Luís Moniz Pereira. Reasoning with Logic Programming, volume 1111 of Lecture Notes in Artificial Intelligence.Springer, 1996.
- [Bra90] Michael Bratman. What is Intention?, in Intentions in Communication. MIT, 1990.
- [CL90a] P. Cohen and H. Levesque. Intention is choice with commitment. Artificial Intelligence, 42(3), 1990.
- [CL90b] Philip Cohen and Hector Levesque. Persistence, Intention, and Commitment, in Intentions in Communication, pages 33--70. MIT, 1990.
- [Car88] Sandra Carberry. Modelling the user's plans and goals. Computational Linguistics, 14(3):23--37, 1988.
- [GS86] Barbara Grosz and Candice Sidner. Attention, intention, and the structure of discourse. Computational Linguistics, 12(3):175--204, 1986.
- [HM87] S. Hanks and D. McDermott. Nonmonotonic logic and temporal projection. Artificial Intelligence, 33, 1987.
- [KR93] 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.
- [KS86] Robert Kowalski and Marek Sergot. A logic-based calculus of events. New Generation Computing, 4:67--95, 1986.
- [LA87] D. Litman and J. Allen. A plan recognition model for subdialogues in conversations. Cognitive Science, (11):163--200, 1987.
- [LA91] Alex Lascarides and Nicholas Asher. Discourse relations and defeasible knowledge. In Proceedings of the 29th Annual Meeting of ACL, pages 55--62, 1991.
- [PR93] J.Pinto and R.Reiter. Temporal reasoning in logic programming: A case for the situation calculus. In D.S. Warren, editor, Proceedings of the 10th ICLP. MIT Press, 1993.
- [Per90] Raymond Perrault. An Application of Default Logic to Speech Act Theory, in Intentions in Communication, chapter 9, pages 161--186. MIT, 1990.
- [Per91] F.Pereira and M.Pollack. Incremental interpretation. Artificial Intelligence, 50:40--82, 1991.
- [Pol90] Martha Pollack. Plans as Complex Mental Attitudes, in Intentions in Communication, chapter 5, pages 77--104. MIT, 1990.
- [QL92] P. Quaresma and J. G. Lopes. A two-headed architecture for intelligent multimedia man-machine interaction. In B. de Boulay and V. Sgurev (eds). Artificial Intelligence V - methodology, systems, applications. North Holland, 1992.
- [QL95] Paulo Quaresma and José Gabriel Lopes. Unified logic programming approach to the abduction of plans and intentions in information-seeking dialogues. Journal of Logic Programming, (54), 1995.
- [RL92] IrenePimenta Rodrigues and José GabrielPereira Lopes. Discourse temporal structure. In Proceedings of the COLING'92, 1992.
- [RL93] IrenePimenta Rodrigues and José Gabriel Lopes. Building the text temporal structure. In Progress in Artificial Intelligence: 6th Portuguese Conference on AI. Springer-Verlag, 1993.
- [RL97] IrenePimenta Rodrigues and José Gabriel Lopes. AI5, An Interval Algebra for the temporal relations conveyed by a text. In Mathematical Linguistics II, Eds Carlos Martin-Vide, John Benjamins, 1997.
- [Son91] F.Song. A Processing Model for Temporal Analysis and its Aplication to Plan Recognition. PhD thesis, University of Waterloo, Waterloo, Ontario, Canada, 1991.