Criação e Acesso a Informação Semântica Aplicada ao Governo Eletrónico

Mário Rodrigues
Universidade de Aveiro
mjfr@ua.pt

Gonçalo Paiva Dias
Universidade de Aveiro
gpd@ua.pt

António Teixeira
Universidade de Aveiro
ajst@ua.pt

Resumo

Os cidadãos, empresas ou serviços públicos - os clientes - que procuram informações no contexto do Governo Eletrónico visam obter respostas objetivas às suas questões. Para isso é necessário que os sistemas de pesquisa consigam manipular a informação de modo a que seja disponibilizada de uma forma eficaz e adequada às necessidades de cada cliente. Uma vez que grande parte dos documentos do governo estão escritos em formatos não estruturados e em linguagem natural, é necessário desenvolver métodos para obter e estruturar este tipo de informação. A alternativa seria indexar pelo seu texto a grande quantidade de documentos existente, uma solução desadequada no contexto do Governo Eletrónico, uma vez que assim seriam retornados frequentemente muitos resultados a cada pesquisa.

Este artigo apresenta um primeiro protótipo de uma aplicação que gera informação semântica a partir de textos escritos em Português. A informação semântica gerada corresponde a um domínio de conhecimento definido por um operador humano através de uma interface gráfica, de modo a que o sistema seja adaptável às diferentes áreas de atuação do Governo Eletrónico. O conteúdo é acessível através de uma interface em linguagem natural e através de uma interface de pesquisa que aceita entradas SPARQL. Deste modo é possível aos clientes aceder diretamente ou integrar este sistema com os seus próprios sistemas de informação. A aplicação está organizada em três grandes módulos: Representação do Conhecimento que permite definir domínio de conhecimento e sua semântica e criar exemplos semente, nos textos, de conceitos do domínio de conhecimento; Processamento de Linguagem Natural que permite obter estruturas sintáticas associadas às frases em linguagem natural; e Extração e Integração Semântica que utiliza os exemplos semente para treinar classificadores estatísticos a identificar nas estruturas sintáticas os conceitos do domínio de conhecimento, que utiliza os classificadores treinados para detetar esses conceitos em estruturas sintáticas de novas frases, e que contém as interfaces para pessoas e máquinas.

Neste artigo apresentamos igualmente exemplos ilustrativos da utilização do sistema e os resultados de uma primeira avaliação de desempenho. O sistema funciona para o Português e foi construído reutilizando software do estado da arte, maioritariamente desenvolvido visando o Inglês. A sua modularidade permite alterar a língua base do sistema, de Português para outra, alterando o módulo de Processamento de Linguagem Natural e sem ser necessário alterar os restantes módulos da aplicação.

1 Introdução

O Governo Eletrónico (e-gov) é uma expressão utilizada para descrever a utilização das Tecnologias da Informação e da Comunicação (TIC) no âmbito do governo e da administração pública. Refere-se a vários conceitos alternativos ou complementares, incluindo o uso das TIC para tornar mais fácil, mais rápido e mais barato o acesso a informação e a serviços aos clientes do governo: cidadãos, empresas, e outros organismos governamentais (Layne e Lee, 2001).

Os órgãos de governo e da administração pública produzem grandes quantidades de informação sob a forma de leis, regulamentos, editais, atas, etc. Estes documentos são normalmente escritos em linguagem natural (Português, Castelhano, etc.) em texto livre, sem uma estrutura em meta linguagem que indique qual o significado das diferentes partes do documento. Mesmo que estes documentos sejam armazenados em computadores, o seu formato de texto livre dificulta a manipulação automática da informação neles contida de modo a ir de encontro às necessidades específicas dos clientes do governo. Frequente são retornados muitos resultados às pesquisas efetuadas com vista a que a informação relevante esteja no conjunto de resposta. Por exemplo, se a procura for "processo de obra Maria", normalmente são devolvidos todos os documentos que contenham (pelo menos) uma das palavras, ordenados pela maior semelhança com a procura, e não apenas aquele(s) que contenha(m) informação acerca de processos de obra aplicados por cidadãs chamadas Maria.

Este comportamento é adequado em sistemas de informação genéricos como os motores de pesquisa da Internet. O mesmo já não acontece quando o contexto é o e-gov. Quando cidadãos, empresas ou serviços públicos procuram informações no contexto do e-gov querem obter respostas às suas questões e não uma lista de documentos acerca de tópicos relacionados. Por outro lado, como o governo tem servir a totalidade da população, incluindo cidadãos com poucos conhecimentos de TIC ou acerca dos processos do governo, as respostas devem ser curtas, claras e concisas de modo a evitar dificuldades na sua localização ou interpretação. Além disso, as respostas devem ser textos criados com base nos documentos oficiais.

É por isso importante desenvolver aplicações e tecnologias que permitam um acesso fácil à informação disponibilizada pelos órgãos de governação e administração. Isso implica a utilização de tecnologias que permitam perceber e manipular o conteúdo de documentos escritos em linguagem natural. O e-gov beneficiaria da existência de sistemas capazes de organizar e integrar diversas fontes de informação e capazes de compreender documentos escritos em linguagens naturais tais como o Português (Rodrigues, Paiva Dias e Teixeira, 2010a).

Em virtude disto temos vindo a desenvolver um sistema que utiliza tecnologias de Processamento de Linguagem Natural (PLN) para interpretar o conteúdo dos documentos, e tecnologias de Representação do Conhecimento (RC) para organizar e manipular o conteúdo obtido. O sistema permite definir os tipos de conteúdo que serão procurados e armazenados, permite a integração de informação relevante de outras fontes, e permite o acesso à informação de diversas formas incluindo perguntas em linguagem natural, por referência geográfica ou através de normas da web semântica.

Focámos a aplicação na disponibilização de informação municipal apesar do sistema poder ser utilizado com diversos tipos de informação. A importância dos municípios reside no fato de serem muitas vezes o ponto mais próximo de serviço para os cidadãos e empresas. São também interessantes devido a integrarem, numa única organização, decisão política e execução administrativa (Paiva Dias, 2006).

Neste artigo apresentamos um sistema capaz de gerar e disponibilizar informação semântica a partir de documentos não estruturados escritos em linguagem natural. A próxima subsecção apresenta trabalho relacionado. A secção 2 descreve detalhadamente a concepção e desenvolvimento do sistema. A secção 3 apresenta os exemplos de utilização e a avaliação de desempenho. O artigo termina, na secção 4, com as respetivas conclusões.

1.1 Trabalho Relacionado

A atividade de investigação em e-gov tem sido geralmente centrada na resolução de problemas como a integração e interoperabilidade de serviços, que são problemas muito importantes e devem continuar a ser estudados. Em tais projetos é geralmente considerado que a informação está no sistema, quer tenha sido colocada manualmente ou utilizando bases de dados existentes, como por exemplo o OneStopGov (Chatzidimitriou e Koumpis, 2008) e o Acess-eGov (Sroga, 2008). Tanto quanto sabemos, até hoje nenhum projeto foi dedicado ao problema da aquisição automática de informações a partir de documentos do governo em linguagem natural, quer para Português quer para outras línguas.

Relativamente à extração de informação, vários projetos foram dedicados à tarefa de extração de informação escalável e independente do domínio. DBPedia (Bizer et al., 2009) é uma base de conhecimento criada pela extração de informação das caixas de informações da Wikipedia e utilizou a estrutura da Wikipedia para inferir a semântica. Uma abordagem semelhante foi seguida para criar a base de conhecimento Yago (Suchanek, Kasneci e Weikum, 2007). Além da Wikipedia, o YAGO também utiliza um conjunto de regras para melhorar a precisão da extracção de informação e a WordNet para desambiguar os significados das palavras. Estas bases de conhecimento foram criados sem qualquer processamento de linguagem natural.

O sistema Kylin (Wu, Hoffmann e Weld, 2008) usa informações das caixas de informação da Wikipedia para treinar classificadores estatísticos que mais tarde são usados para extrair informações a partir de textos de linguagem natural. Os textos são analisados pelas suas etiquetas morfo-sintáticas e caraterísticas de superfície (posição das palavras na frase, a capitalização, a presença de dígitos ou carateres especiais, etc.) Não usa informação sintática.

Outros sistemas do estado da arte não utilizam a Wikipedia como fonte de conhecimento. O TextRunner (Banko et al., 2007) pretende extrair todas as instâncias de todas as relações significativas a partir de páginas web. Constrói a sua ontologia a partir do corpus sem controlar se as relações ontológicas estão bem definidas e sem desambiguar as entidades. O KnowItAll (Etzioni et al., 2004) utiliza exemplos especificados manu-

almente que expressam um conjunto de relações, por exemplo `amigo(João,Pedro)`. Esses exemplos são utilizados para obter padrões textuais que podem expressar as relações, por exemplo "o João é amigo do Pedro". Os padrões textuais são usados para treinar um conjunto de informações pré-definidas.

O sistema Leila (Suchanek, Ifrim e Weikum, 2006) aperfeiçoou o método do KnowItAll usando tanto exemplos e contra-exemplos como sementes, a fim de gerar padrões mais robustos, e usando análise sintática para gerar padrões de extração de informações. A maior robustez dos padrões conjugada com uma análise sintática que permite capturar informações em frases mais complexas foram as pricipais razões para que esta abordagem fosse adoptada no nosso sistema.

Relativamente a interfaces em linguagem natural escrita o que tem sido estudado é, essencialmente, como mapear frases em linguagem natural para os esquemas de armazenamento de informação. Um dos sistemas relevantes é o NaLIX, uma interface para bases de dados XML que aceita frases arbitrárias em Inglês. Esta interface traduz as pesquisas em expressões XQuery e, por exemplo, é possível consultar uma base de dados acerca de filmes com frases do tipo "find the title of publications with more that 5 authors" que traduz para: encontra os títulos de obras com mais de 5 autores (Li, Yang e Jagadish, 2005).

O Panto é outra interface em linguagem natural escrita que aceita consultas genéricas em linguagem natural, produzindo como saída consultas *Simple Protocol and RDF Query Language* (SPARQL), que é atualmente a linguagem padrão para acesso de dados da Web semântica. Foi concebido para ser aplicável a qualquer ontologia não pressupõe nada acerca do domínio do conhecimento. Os seus autores argumentam que obtém bons resultados e que ajuda a fazer a ponte entre a lógica da web semântica e os utilizadores (Wang et al., 2007).

O ESTER é um sistema modular que conjuga pesquisas de texto completo e pesquisas em ontologia. Responde a consultas SPARQL básicas reduzindo-as a um pequeno número de duas operações básicas: pesquisa de prefixo e junção. Suporta uma mistura de consultas semânticas com consultas de texto normais e sugere ao utilizador possíveis interpretações semânticas da consulta (Bast et al., 2007).

2 Sistema Desenvolvido

O sistema desenvolvido está organizado conforme o modelo conceptual apresentado na Figura 1. O modelo separa claramente o domínio da linguagem natural do domínio da representação do conhecimento e está organizado em três componentes:

- Representação do Conhecimento - componente que contém ferramentas para definir a semântica do sistema - através de uma ontologia representada em *Web Ontology Language Description Logic* (OWL-DL) - e para permitir a operadores humanos adicionar exemplos de correspondência entre essa semântica e elementos presentes nos textos;
- Processamento de Linguagem Natural - componente baseado em tecnologias da área de PLN que inclui tecnologias de processamento de informação para obter estruturas sintáticas que representam as frases encontradas nos textos em linguagem natural. Conforme os exemplos definidos na RC, algumas destas estruturas serão associadas a elementos da ontologia;
- Extração e Integração Semântica - componente que aprende as associações entre as estruturas sintáticas e a ontologia e aplica-as a novos textos para obter novas informações. Este componente pode complementar a informação contida nos textos com fontes estruturadas de informação, como por exemplo coordenadas geográficas dos locais via Google Maps API e organização política do território via Geo-Net-PT01 (Chaves, Silva e Martins, 2005). Inclui ainda interfaces de acesso aos dados.

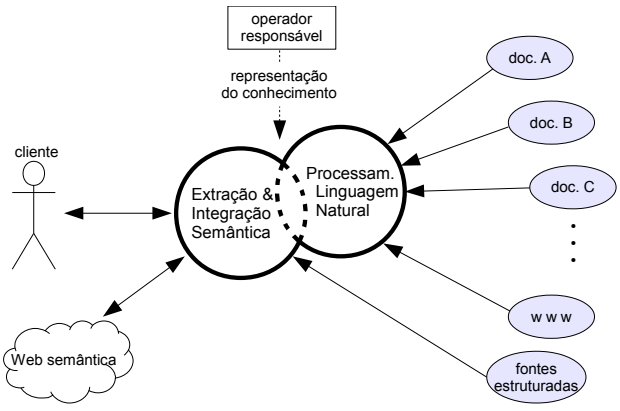

Figura 1: O modelo conceptual. A informação semântica é extraída das estruturas provenientes do PLN conforme definido pela RC definida pelo operador responsável pelo sistema. As setas unidirecionais representam aquisição do conteúdo e as bidirecionais representam as interfaces.

O resultado é informação semântica que pode ser consultada e acedida em vez, ou em complemento, dos documentos originais (ver Figura 1).

O sistema foi construído reutilizando software de código aberto - algum adaptado para trabalhar com Português - para tirar vantagem do estado da arte em termos de abordagens e ferramentas existentes. Foi desenvolvido software especifico para integrar o software reutilizado num sistema coerente. A arquitetura da instanciação do modelo conceptual está representada na Figura 2 e será descrita em mais detalhe nas subsecções que se seguem.

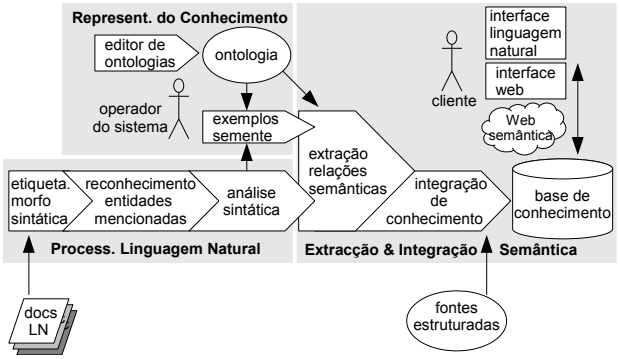

Figura 2: Instanciação do modelo conceptual. Os três grandes módulos são delimitados pelo sombreado. A Representação do Conhecimento define a semântica do sistema e fornece exemplos semente dos conceitos nos textos. O módulo de Processamento de Linguagem Natural enriquece o texto com etiquetas morfo-sintáticas, entidades mencionadas e estruturas sintáticas. O módulo de Extração e Integração Semântica treina modelos de extracção com base nos exemplos semente, aplica-os em todos os textos, integra outras fontes de informação e disponibiliza a informação aos clientes.

2.1 Representação do Conhecimento

O primeiro passo para construir uma representação do conhecimento é definir uma estrutura que represente conceitos de um domínio e respetivas relações. Para isso utilizámos ontologias que formalmente são definidas como "*a formal, explicit specification of a shared conceptualisation*", o que traduz para: especificação explícita e formal de uma conceptualização partilhada (Gruber, 1993). Ser "explícita" implica que todos os conceitos usados e respetivas restrições têm de estar definidos explicitamente e ser "formal" refere-se a ter de ser legível para máquinas. Uma "conceptualização" é um modelo abstrato de que representa um domínio, identificando conceitos e relações relevantes a essa parte do mundo. Ser "partilhada" é importante porque uma ontologia deverá servir para partilhar conhecimento e por isso deve ser aceite por um grupo ao invés de ficar restrita a um indivíduo.

As ontologias permitem representar um conjunto de conceitos pertencentes a um domínio, bem como as relações existentes entre esses conceitos. O fato de ser uma especificação formal bem definida permitiu o desenvolvimento de ferramentas de *software* que inferem novos fatos através de implicações lógicas acerca dos dados já conhecidos. Na nossa aplicação as ontologias são criadas e/ou editadas usando o Protégé (versão 4). Para o domínio do e-gov, a ontologia criada inclui as ontologias Friend-of-a-Friend (FOAF) (Brickley e Miller, 2010), Dublin Core (Weibel et al., 2007), World Geodetic System revisão de 1984 (National Imagery and Mapping Agency, 2000), e GeoNames versão integral (GeoNames, 2010). Inclui também classes especificamente criadas para lidar com assuntos relativos aos municípios. Foi criada uma classe denominada *Assunto_executivo* que é subclasse da classe de nível superior *Thing* e que possui sete subclasses (ver Figura 3). As subclasses de *Assunto_executivo* e respetiva descrição encontram-se na Tabela 1.

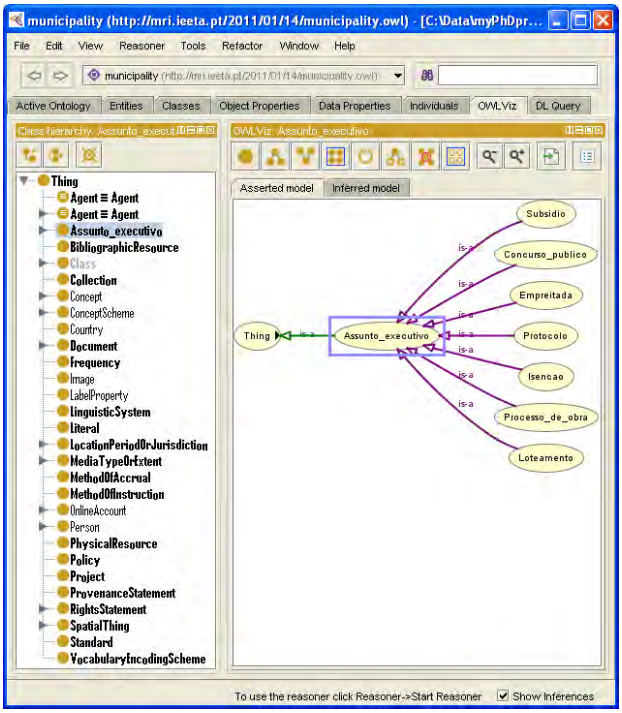

Figura 3: A interface de criação da ontologia. No painel da esquerda está a lista de todas as subclasses de *Thing*. No painel direita encontra-se uma representação gráfica da classe *Assunto_executivo* e respetivas subclasses.

Cada assunto executivo pode conter seis propriedades para estabelecer relações com outras classes da ontologia, como por exemplo com as classes de *Person* (Pessoa) das ontologias importadas FOAF e Dublin Core. As propriedades estão enumeradas e descritas na Tabela 2.

Classe	Descrição
Loteamento	Pedido de permissão para lotear ou alterar loteamentos de terrenos.
Empreitada	Relativo a processos de construção em execução.
Processo_de_obra	Anúncios relativos a processos de construção genéricos: início de trabalhos, alterações em orçamentos, expropriações, etc.
Isenção	Pedidos de isenções de taxas e outros pagamentos municipais.
Protocolo	Protocolos assinados com outras instituições.
Concurso_publico	Anúncios de concursos públicos relativos a aquisição de equipamento, construções, contratação, etc.
Subsidio	Subsídios pedidos e/ou concedidos pela autarquia.

Tabela 1: Subclasses da classe *Assunto_executivo* e respetiva descrição.

Propriedade	Descrição
deliberação	Resultado do pedido.
identificador	Identificador unívoco dado pelos serviços municipais.
montante	Qualquer quantia de dinheiro envolvida no processo.
motivo	O motivo do processo.
local	O local da construção ou do loteamento, morada da entidade que assinou o protocolo ou que pediu isenção ou subsídio.
submetidoPor	Entidade ou entidades que estão envolvidas no processo, excluindo o município.

Tabela 2: Tipos de relações associadas à classe *Assunto_executivo*.

2.1.1 Exemplos Semente

Após a definição do domínio do conhecimento é necessário encontrar exemplos dos conceitos nos textos que o sistema deverá processar. Estes exemplos serão utilizados para treinar algoritmos de aprendizagem automática de modo a que o sistema detete esses conceitos em todos os documentos a processar.

A associação entre as amostras de texto e classes da ontologia e as relações são feitas usando o anotador AKTive Media (Chakravarthy, Ciravegna e Lanfranchi, 2006). No arranque da interface de anotação é necessário escolher ou criar uma sessão de anotação e escolher os textos a anotar e a ontologia que define o domínio do conhecimento. Após este passo é possível iniciar o processo de anotação ou então pedir ao sistema para pré-anotar partes do texto.

A pré-anotação foi uma funcionalidade desenvolvida para facilitar o processo de anotação quando existe uma grande quantidade de textos a anotar. Serve para pré-anotar no texto as classes da ontologia mas não as relações da ontologia. O seu comportamento é definido por um ficheiro de configuração que contém, em cada linha, uma entrada com uma expressão regular a localizar seguida da classe ou classes da ontologia a associar a essa palavra (ver Figura 4).

As palavras pré-anotadas ficam destacadas por um fundo colorido em que a cor está associada com a classe da ontologia (ver Figura 5). No fim do processo de anotação todas as pré-anotações que não foram validadas ou completadas pelo utilizador serão descartadas. Deste modo os exemplos semente são todas e apenas as anotações validadas pelo utilizador.

Figura 4: Ficheiro de configuração da pré-anotação. Cada linha contém a expressão regular a detetar no texto seguida da(s) classe(s) da ontologia a associar ao texto abrangido por essa expressão regular.

O procedimento para anotar uma frase é o seguinte (ver Figura 5):

1. Selecionar a classe da ontologia no painel superior esquerdo. Ao escolher a classe da ontologia surgem, no painel debaixo da caixa de procura, as relações possíveis para essa classe;
2. Selecionar a(s) palavra(s) a associar a essa classe. As palavras ficam por cima de um fundo colorido cuja cor está associada à classe escolhida;
3. Escolher a relação da ontologia a associar ao texto selecionado e marcar no texto o objeto dessa relação. A relação surgirá no painel inferior esquerdo;
4. Repetir o passo 3 até todas as relações estarem marcadas;
5. Voltar ao passo 1 até todo o texto relevante estar marcado.

A Figura 5 mostra a anotação de um subsídio cuja motivação é "execução do Plano Anual e a Escola Artística", foi submetido pela "ARCEL" e o montante envolvido é "8.640,00€".

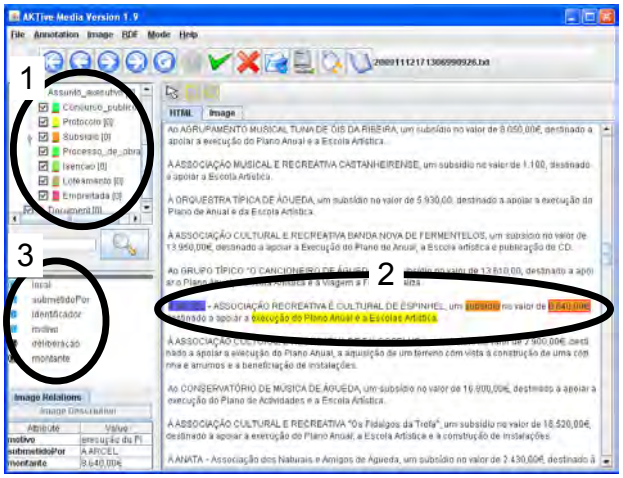

Figura 5: A interface de anotação. Os números correspondem à sequência de passos descritos no procedimento.

O resultado desta etapa é a ontologia e o conjunto de exemplos nos textos de entidades das classes e relações da ontologia.

2.2 Processamento de Linguagem Natural

Esta parte do sistema inclui ferramentas para obter e extrair o conteúdo de documentos da Web e/ou do sistema de ficheiros local. O sistema permite definir a fonte dos dados, sendo o conteúdo dos ficheiros processado automaticamente num encadeamento de operações sem intervenção dos utilizadores. A sequência de operações é igual à encontrada num vasto conjunto de sistemas de PLN (um bom exemplo é (Ferreira et al., 2009)): etiquetagem morfo-sintática, reconhecimento e classificação de entidades mencionadas e análise sintática.

O primeiro passo, a etiquetagem morfo-sintática (em Inglês *Part of Speech (POS) tagging*), tem por objetivo associar os diversos elementos do texto com classes morfo-sintáticas tais como substantivo, adjetivo, etc (Mihalcea, 2010). No sistema implementado a etiquetagem é realizada pelo TreeTagger que anota o texto com etiquetas morfo-sintáticas e com lemas e tem sido usado com sucesso para marcar várias linguagens naturais, incluindo Português (Schmid, 1994). O TreeTagger foi treinado com o Bosque v7.3, uma versão especificamente escolhidas por ser a única no formato aceite também pelo analisador sintático (descrito adiante). O Bosque é um subconjunto da Floresta (Freitas, Rocha e Bick, 2008) revisto por linguistas. O léxico utilizado foi enriquecido com o LABEL-LEX-sw (Ranchhod, Mota e Baptista, 1999).

De seguida é efetuado o Reconhecimento de Entidades Mencionadas (REM) e respetiva classificação que tem por objetivo detetar e classificar elementos atómicos no texto em categorias pré-definidas tais como nomes de pessoas, organizações, locais, etc (Santos e Cardoso, 2007). Além das classes de REM e seu significado serem diferentes das de POS *tagging*, uma diferença fundamental é que o processo de REM implica frequentemente o agrupamento de palavras numa única entidade. O REM do sistema é feito com o Rembrandt (Cardoso, 2008). O Rembrandt é um sistema de REM desenvolvido para Português que utiliza a estrutura e conteúdo da Wikipédia como uma fonte de conhecimento para classificar todos os tipos de entidades mencionadas no texto. Rembrandt tenta classificar cada entidade mencionada de acordo com as diretivas do segundo HAREM (Mota e Santos, 2008).

O terceiro passo, a análise sintática, é o processo de determinar a estrutura gramatical de uma sequência de palavras segundo uma determinada gramática formal. A análise sintática transforma um texto numa estrutura de dados. Este passo é efetuado por um analisador, em Inglês *parser*, de dependências chamado MaltParser (Hall et al., 2007). O MaltParser já foi utilizado com sucesso para analisar várias línguas o Inglês, Francês, Grego, Sueco e Turco. Foi treinado para Português com o Bosque v7.3 que existe no formato aceite por esta ferramenta, o formato CoNLL-X.

O funcionamento geral deste módulo está esquematizado na Figura 6.

2.3 Extração e Integração de Informação Semântica

Esta parte do sistema tem dois modos de operação: modo de treino e modo de execução.

No modo de treino, o sistema aprende a associar as estruturas sintáticas das frases às classes e relações da ontologia. Esta aprendizagem é baseada em exemplos anotados manualmente.

No modo de execução, o sistema aplica as associações aprendidas às estruturas sintáticas de todas as frases dos documentos a processar para extrair classes e relações semânticas do texto. O procedimento de ambos os modos é explicado seguidamente.

2.3.1 Treino de Modelos para Extração

Resumidamente, o processo desenrola-se da seguinte forma:

1. Processar todos os documentos de treino com o módulo de PLN para se obter estruturas sintáticas de todas as frases de treino;

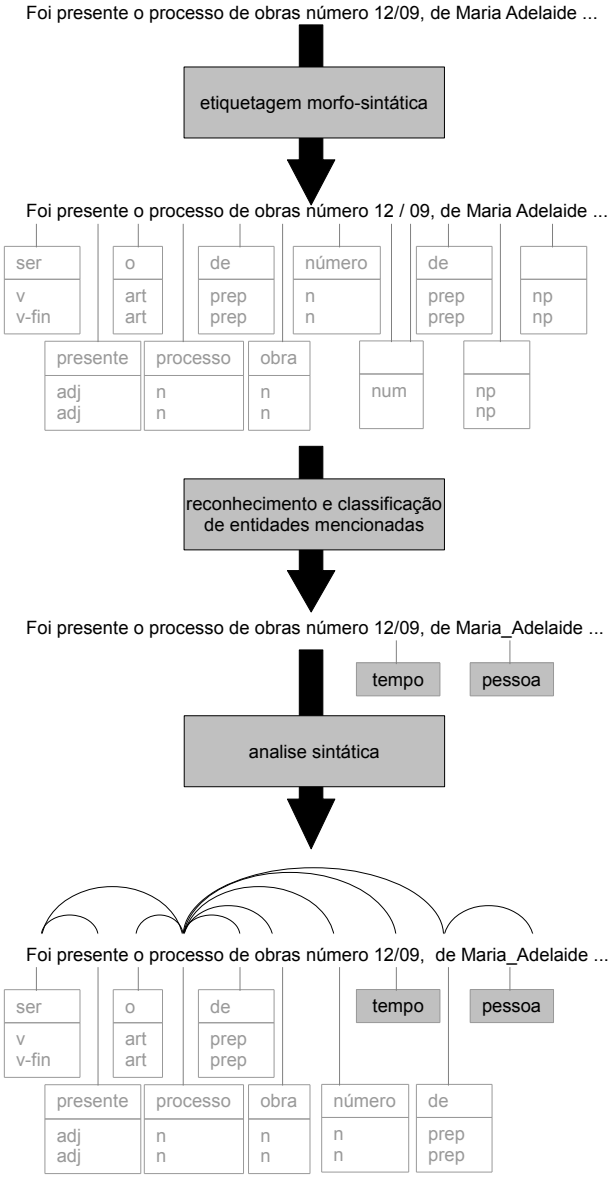

Figura 6: Sequência de passos do Processamento de Linguagem Natural e respetivos resultados intermédios. A entrada do módulo são frases sem estrutura definida e a saída é uma estrutura sintática enriquecida com etiquetas morfo-sintáticas e entidades mencionadas.

2. Para todas as estruturas sintáticas do conjunto de treino e para todas as relações (semânticas) anotadas: se as palavras da relação estiverem na estrutura sintática fazer os passos 3 e 4;
3. Guardar o caminho da árvore sintática entre as palavras envolvidas na relação. Este caminho é considerado um exemplo de elemento da relação ontológica e é composto por: sequência de ligações necessárias e lemas e etiquetas morfo-sintáticas das palavras que estão no caminho;
4. Guardar o contexto das palavras envolvidas na relação. O contexto é considerado um exemplo de elemento da classe ontológica e é composto por: lema, etiqueta morfo-sintática e tipos ligações sintáticas que a palavra possui;

No final deste processo, os caminhos encontrados para cada relação da ontologia são agrupados e utilizados para gerar um classificador estatístico por relação. Também os contextos encontrados para cada classe da ontologia são agrupados e utilizados para gerar um classificador estatístico por classe. Os classificadores estatísticos utilizados são baseados no algoritmo *k-nearest neighbor* (k-NN) e são semelhantes aos utilizados no LEILA (Rodrigues, Paiva Dias e Teixeira, 2011).

O último passo é melhorar a precisão dos classificadores. Este passo assume que todas as relações existentes no conjunto de treino foram marcadas. Assim, os classificadores começam por avaliar todas as estruturas sintáticas do conjunto de treino. Todas as estruturas sintáticas que são avaliadas como representando classes e relações da ontologia e que não as representam, ou seja não são exemplos anotados, passam a ser contra-exemplos para o classificador que gerou essa avaliação errada. Após da recolha de todos os contra-exemplos, todos os classificadores são novamente treinados agora utilizando os exemplos e os contra-exemplos.

2.3.2 Aplicação dos Modelos

À semelhança do treino, a execução da extração de informação inicia-se com o módulo de PLN a processar todos os documentos de modo a se obterem estruturas sintáticas para todas as frases dos documentos. Seguidamente, os classificadores estatísticos gerados na fase de treino avaliam se as estruturas sintáticas representam alguma classe ou relação da ontologia. Caso a avaliação do classificador seja mais elevada que o limiar de aceitação, essa informação é recolhida para uma base de conhecimento temporária.

Após a extração de informação segue a integração de informação. O motor de inferência semântico aplica as regras ontológicas a todos os dados e verifica se não existem implicações impossíveis, ou seja verifica se a nova informação é coerente com a ontologia e com informação já presente no sistema. O motor de inferência utilizado é o Pellet (Sirin et al., 2007) que suporta integralmente o formalismo OWL-DL. Toda a informação coerente passa da base de conhecimento temporária para a base de conhecimento do sistema. A informação incoerente não é adicionada e gera um aviso no registo do sistema para se averiguar a causa da incoerência. A base de conheci-

mento é armazenada e gerida pelo Virtuoso Universal Server[1]. Este servidor tem, entre outras características, um motor de base de dados nativo para *Resource Description Language* (RDF), suporta pesquisas SPARQL e, como indicador do seu desempenho, é o servidor da DBpedia[2] que contém atualmente 3,64 milhões de fatos dos quais 1,83 milhões estão classificados numa ontologia consistente (416.000 pessoas, 526.000 lugares, 169.000 organizações, etc.).

Nesta fase também se procura informação em falta, de acordo com a ontologia, em fontes externas estruturadas de informação. É necessário que a informação proveniente destas fontes seja estruturada de modo a se poder definir uma semântica apropriada para elas, uma vez que nesta fase do processamento a informação entra diretamente na base de conhecimento, não passando pelos classificadores estatísticos responsáveis por detetar informação semântica relevante.

Por agora existem dois tipos de informação adicionados caso estejam em falta na base de conhecimento: as coordenadas *Global Positioning System* (GPS) de entidades que deverão ter uma localização fixa e a organização política dos espaços.

As entidades que estão definidas na ontologia como tento uma localização fixa são, por exemplo, cidades, ruas, sedes de organizações e alguns eventos. Nestes casos, caso não existam na base de conhecimento, as coordenadas GPS destes locais são consultadas via Google Maps API.

Sendo esta aplicação um sistema de pesquisa de informação para a área do e-gov é relevante saber quais os locais políticos relativos à informação (rua $\subset$ freguesia $\subset$ cidade $\subset$ concelho...). Assim, além das coordenadas GPS também é adicionada a organização política dos espaços que é obtida utilizando uma ontologia geográfica de Portugal com cerca de 418 mil entradas chamado Geo-Net-PT01 (Chaves, Silva e Martins, 2005).

Estes dois tipos de informação adicionados permitem o sistema exibir informações espacialmente num mapa e procurar e relacionar informações em função da sua localização (Rodrigues, Paiva Dias e Teixeira, 2010b).

2.4 Interfaces de Acesso à Informação

Foram implementadas duas formas de aceder à informação gerida pelo sistema. Uma destina-se a ser utilizada de um modo fácil e intuitivo por pessoas e corresponde ao acesso via interface de linguagem natural. A outra destina-se a ser utilizada por sistemas que queiram aceder à informação semântica contida na base de conhecimento e corresponde ao acesso via interface para máquinas. Ambas as interfaces são explicadas seguidamente.

2.4.1 Interface para Utilizadores Humanos

A interface para humanos suporta linguagem natural escrita e permite a interacção usando Português. É uma interface flexível o suficiente para permitir a pesquisa por palavras chave, tal como os motores de pesquisa da Web, ou através da formulação de perguntas em Português. Esta flexibilidade é importante uma vez que o e-gov tem de servir a totalidade da população independentemente do seu nível de proficiência nas TIC. Assim, utilizadores habituados a pesquisar informação na Web podem pesquisar de um modo que já lhes é familiar ou então podem formular as perguntas às quais querem obter respostas.

A interface utilizada é baseada no NLP-Reduce (Kaufmann e Bernstein, 2007), uma interface em linguagem natural para a web semântica, em Inglês e independente do domínio. A escolha do NLP-Reduce foi motivada por esta independência de domínio, o que o torna adaptável aos vários assuntos do e-gov, e por ser facilmente adaptável ao Português uma vez que não contém componentes específicos para processar Inglês. A sua abordagem evita deliberadamente quaisquer tecnologias semântica ou linguística complexas e não interpreta ou tenta compreender as perguntas efetuadas. Consiste em associar as palavras (e seus sinónimos) contidas na pergunta às expressões utilizadas para descrever classes, relações e indivíduos presentes na base de conhecimento. Deste modo, se a ontologia estiver descrita Português, uma parte considerável do sistema fica automaticamente em Português. Apenas foram necessárias pequenas adaptações para Português na formulação das perguntas como por exemplo palavras muito frequentes e com pouco significado (*stopwords*) e os pronomes interrogativos (qual, quem, etc.).

A interface constrói automaticamente um léxico usando as palavras contidas em todos os fatos explícitos ou inferidos da base de conhecimento. Ao léxico são igualmente adicionados os sinónimos das palavras já presentes nele. A procura de sinónimos é efetuada através da ontologia lexical PAPEL (Oliveira, Santos e Gomes, 2010), criado pela Linguateca a partir do Dicionário PRO da Língua Portuguesa da Porto Editora. Também são adicionadas ao léxico os lemas das palavras nele presentes de modo a aumentar a abrangência lexical.

[1] http://virtuoso.openlinksw.com/
[2] http://dbpedia.org/

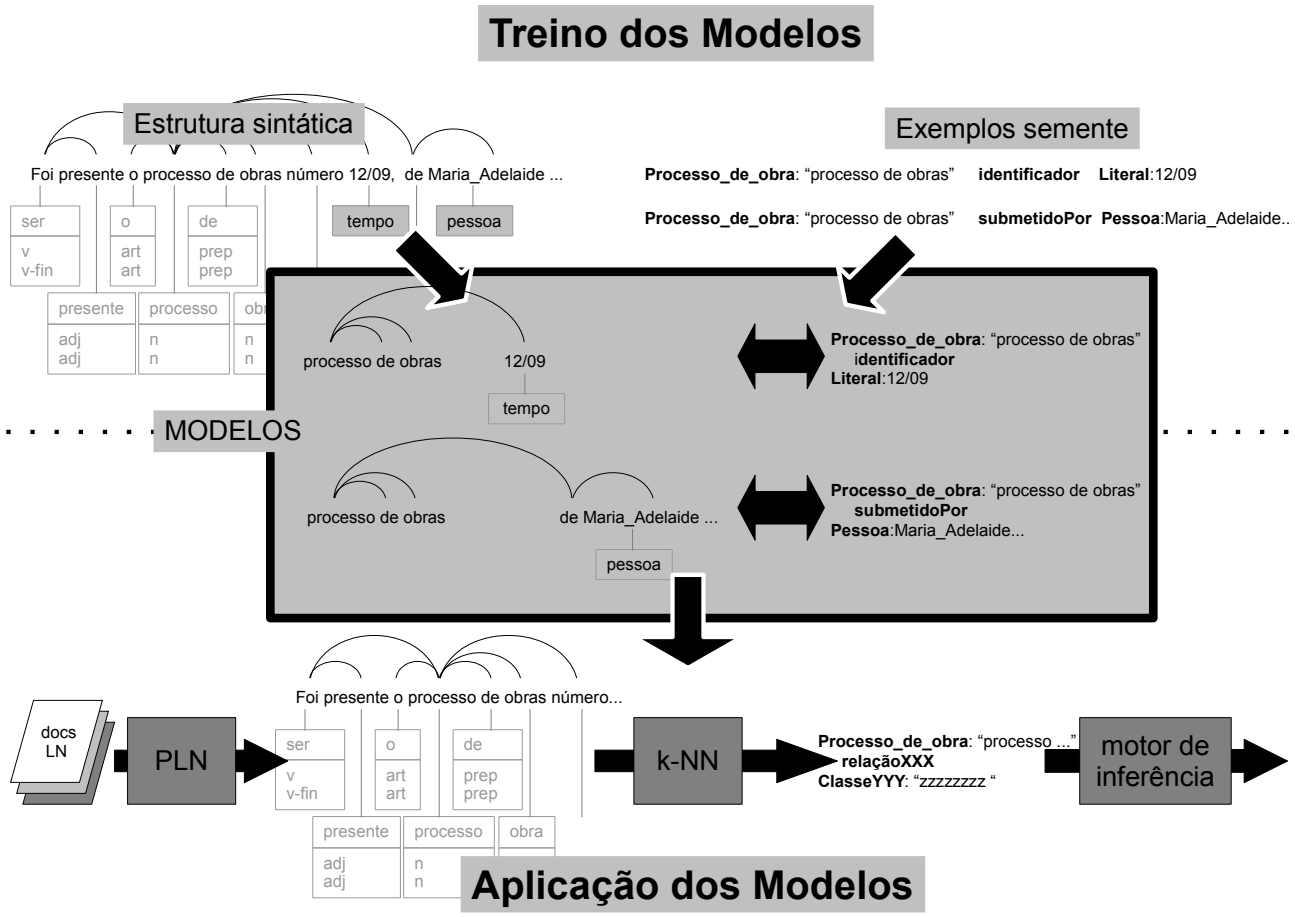

Figura 7: Treino dos modelos e respetiva aplicação. O treino começa por guardar o caminho da árvore sintática que liga as palavras envolvidas no exemplo semente. Após a recolha de todos os exemplos (e contra-exemplos) os caminhos são usados para treinar um classificador estatístico baseado no algoritmo k-NN. Durante a execução, os classificadores treinados são aplicados avaliam todas as estruturas sintáticas de modo a verificar se estas representam uma relação ontológica.

O processamento das entradas dos utilizadores inicia-se com a remoção de sinais de pontuação e de *stopwords*, tais como artigos, preposições, algumas conjunções. Seguidamente, com base nos lemas das palavras sobrantes, é construído uma pesquisa SPARQL que será submetida a um motor de pesquisas SPARQL. Considere-se o exemplo da pergunta "Qual a deliberação do processo de obra submetido por Maria?". A construção da pesquisa SPARQL é efetuada do seguinte modo:

1. São procurados os fatos em que pelo menos um dos lemas da pesquisa faz parte da etiqueta de uma propriedade de objeto. Considerando o exemplo, os fatos contendo as propriedades <deliberacao> e <submetidoPor> serão retornados. As propriedades de objeto encontradas são ordenadas de acordo com o ajustamento entre a sua etiqueta e as palavras da pesquisa, por exemplo a etiqueta <submetidoPor> obtém melhor classificação com as palavras "submetido por" que uma etiqueta que fosse <submetido>;
2. São procuradas no léxico elementos que podem ser conjugados com as propriedades encontradas no passo 1, usando os restantes lemas da pesquisa e tomando em consideração os seus domínio e contra-domínio. No nosso exemplo são procuradas os elementos que contenham "qual", "processo", "obra" e "maria". Como a classe <Processo_de_obra> contém a palavra "processo" e "obra" e é o domínio de ambas as propriedades obtidas no passo 1, este passa a ser o elemento de ligação entre elas;
3. São procuradas no léxico as propriedades relativas a dados cujos valores correspondem aos restantes lemas da pesquisa. Estas propriedades são combinadas com as identificadas anteriormente, tendo em conta os domínios e contra-domínios de todas as propriedades envolvidas e ordenados conforme o seu ajustamento às palavras sobrantes. Das palavras sobrantes do nosso exemplo, "qual"

e "maria", a palavra "maria" existe como valor da propriedade <Nome>. Como o domínio de <Nome> é a classe <Pessoa> que por sua vez é contra-domínio da relação <submetidoPor>, a propriedade <Nome> é adicionada à pesquisa.

4. Por último é gerada a pesquisa SPARQL com a junção de propriedades que obtiveram a classificação mais alta nos passos 1 e 3. Adicionalmente são removidos os duplicados semanticamente equivalentes e é efetuada a pesquisa com o SPARQL gerado.

2.4.2 Interface para Máquinas

A interface para máquinas aceita como entrada pesquisas em SPARQL e devolve um RDF contendo o conjunto de resultados e respetiva marcação semântica. Esta interface pode ser utilizada pela interface em linguagem natural, depois de gerar a pesquisa SPARQL, ou por sistemas externos que pretendam aceder à informação semântica. O seu objetivo é possibilitar a interoperabilidade entre este e outros sistemas.

A interoperabilidade é importante para o conceito de e-gov como uma plataforma. Este conceito é uma visão para o futuro em que um dos papéis principais dos sistemas de e-gov é o fornecimento de informação usando formatos abertos e livres e interpretáveis por máquinas (Frissen et al., 2007; United Nations, 2010). A ideia é que se a informação estiver disponível, existirá maior transparência na definição de políticas públicas e permitirá que entidades extra-governamentais utilizem essa informação combinando-a de formas inovadoras e úteis para as populações.

A interoperabilidade também tem um papel central na Web semântica, um conceito introduzido em (Berners-Lee, Hendler e Lassila, 2001). A Web semântica é uma extensão da Web atual que visa atribuir um significado aos conteúdos de modo que seja perceptível por pessoas e por computadores simultaneamente. Uma vez que a ontologia é tornada pública e o seu modo de acesso é uma norma aberta, qualquer entidade externa tem conhecimento do tipo de dados contidos na base de conhecimento, do seu significado semântico e de qual o protocolo de acesso. Uma forma de explorar esta funcionalidade é mostrada na Secção 3.

3 Exemplos de Utilização

As experiências relatadas nesta secção foram concebidas para extrair informações sobre os assuntos municipais públicos mais frequentes e mais procurados por cidadãos e empresas. Para isso foram selecionados três temas em atas municipais públicas: os subsídios concedidos, as licenças de construção solicitadas, e protocolos assinados com outras instituições.

Um *crawler* web obteve todos os documentos disponíveis nos portais da Internet de sete municípios portugueses. Foram selecionados dois conjuntos aleatórios e disjuntos de 50 documentos cada. O documentos selecionados estavam no formato pdf. Um conjunto foi anotado manualmente por uma pessoa e as anotações foram utilizadas para treinar o sistema de de classificação. O outro conjunto foi utilizado em tempo de execução para ter conhecimento extraído pelo sistema.

Os utilizadores podem obter informação utilizando a interface de linguagem natural. A Figura 8 apresenta uma captura de ecrã contendo a resposta à pergunta "Qual a deliberação do processo de obra submetido por Maria?". Na janela por baixo da pergunta verifica-se que o SPARQL gerado é (URL's e variáveis SPARQL abreviados para ficar mais conciso):

```
select distinct * WHERE {
 ?Proc <#SubmetidoPor> ?Pess .
 ?Proc <#Deliberacao> ?Delib .
 ?Pess <#Nome> ?Pess_Nome .
 FILTER(REGEX(?Pess_Nome,'maria','i')).
 ?Proc <#type> <#Processo_de_obra> .
 ?Pess <#type> <#Pessoa>
}
```

A resposta à interrogação SPARQL gerada contém apenas duas entradas na base de conhecimento. A vantagem de ter uma base de conhecimento semântica fica patente neste exemplo uma vez que o sistema associa as palavras "processo de obra" à classe da ontologia "Processo_de_obra", associa a palavra "Maria" a uma pessoa, e procura obter o valor da propriedade "deliberacao". Assim apenas são verificadas as informações que o sistema capturou como relacionando processos de obra com pessoas chamadas Maria e não todas as frases que incluem (algumas das) palavras "processo", "obra" e "Maria". Outra vantagem é ser possível mostrar imediatamente apenas as informações consideradas relevantes, tais como o resultado da deliberação e o nome completo da pessoa, sem mostrar todos os outros dados conhecidos. Contudo é possível obter mais dados uma vez que também são devolvidas as referências da base de conhecimento correspondentes aos processos de obra retornados.

A funcionalidade do acesso para aplicações externas em SPARQL é demonstrada com uma página web (Figura 9) onde são mostrados num mapa os locais que estão envolvidos nos

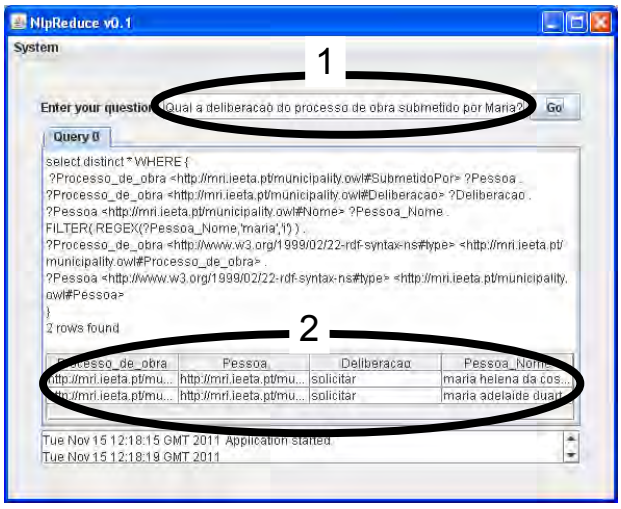

Figura 8: A interface de linguagem natural escrita. A entrada da pergunta é efectuada em 1 e a resposta dada em 2. Neste caso, solicitar significa que a Câmara Municipal solicitou mais documentação. Entre 1 e 2 pode-se ver a pesquisa SPARQL gerada.

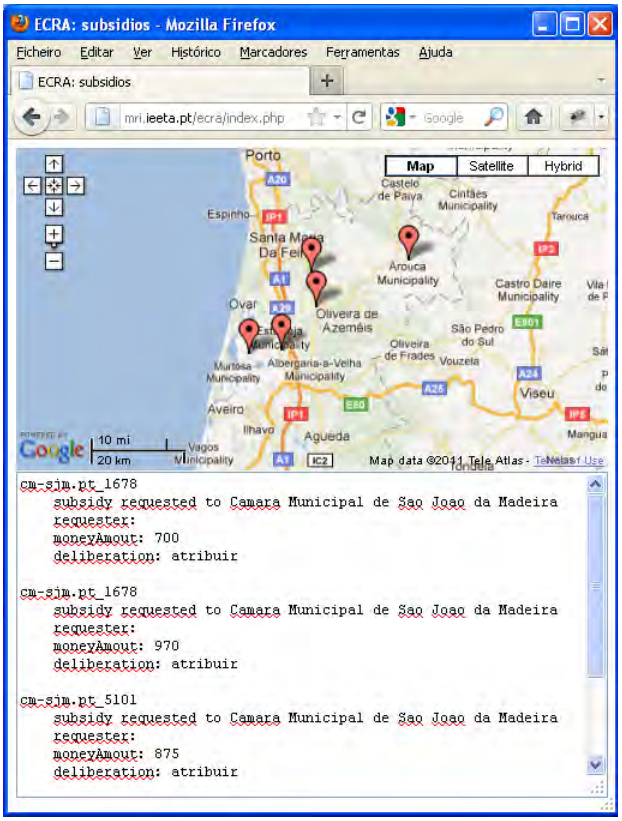

Figura 9: A interface Web. No mapa estão marcados os locais para os quais existe informação. Ao pressionar um local são mostradas as informações relativas ao mesmo na parte inferior da página.

subsídios existentes na base de conhecimento do sistema. A página web desenvolvida faz uma pergunta SPARQL onde questiona várias informações como a latitude e longitude das entidades que concederam os subsídios, o montante de dinheiro pedido e se foi atribuído e a quem. Depois de obter a resposta em RDF, a página web exibe a informação num mapa, usando para isso as coordenadas de latitude e longitude obtidas. Ao selecionar uma localização são mostradas todas as informações relativas a essa localização.

3.1 Avaliação de Desempenho

Recentemente foi efetuada uma avaliação de desempenho do sistema e os resultados obtidos foram apresentados na EPIA2011 - 15th Portuguese Conference on Artificial Intelligence (Rodrigues, Paiva Dias e Teixeira, 2011). A avaliação implicou que uma pessoa verificasse que fatos relevantes estavam contidos nos documentos do conjunto de teste. O conjunto detetado pela pessoa passou a ser a "verdade" e serviu de base de comparação para verificar que fatos foram encontrados ou não pelo sistema, e quais os que foram incorretamente extraídos. Os fatos foram considerados detetados se o sistema extraiu o tipo de fato (subsídio, processo de obra, protocolo) mesmo que estivessem em falta alguns dados como os pretendentes e as quantias envolvidas. Os resultados estão sumarizados na Tabela 3.

Existiam um total de 32 subsídios no conjunto de teste, dos quais o sistema detetou 14 e não houve nenhum falso positivo, isto é, todos os subsídio detetados eram realmente subsídios. Relativamente a processos de obra o sistema detetou 67 de um total de 68. Contudo foram também extraídos como processos de obra 4 informações que não o eram: 4 falsos positivos, Quanto aos protocolos, o sistema detetou 8 dos 41 existentes e houve um falso positivo. A baixa cobertura na deteção de protocolos (0.20) está em grande parte associada à existência de enumerações. Uma vez que estas não existiam no conjunto de treino, o sistema apenas detetou a primeira instituição em enumerações do tipo "... protocolos ... com as seguintes instituições:" seguida da listagem de instituições, uma por linha. Esta falha causou a baixa cobertura uma vez que por cada instituição listada, à exceção da primeira, foi considerado um protocolo não identificado.

O desempenho global do sistema relativamente à extração de informação semântica (precisão 0.95; cobertura 0.63) está em linha com o estado da arte para Inglês: DBpedia (precisão 0.86 a 0.99; cobertura 0.41 to 0.77), Kylin (precisão 0.74 a 0.97; cobertura 0.61 a 0.96), e YAGO/NAGA (precisão 0.91 a 0.99; cobertura

	município									
	a	b	c	d	e	f	g	precisão	cobertura	F_1
subsídio	0(2)	3(3)	4(11)	1(1)	1(1)	3(14)	0(0)	1.00	0.44	0.61
processo de obra	3(4)[1]	13(13)	47(47)	0(0)	0(0)	4(4)	0(0)	0.94	0.99	0.97
protocolo	3(4)	3(3)	0(3)	0(0)	7(24)	2(7)[2]	0(0)	0.89	0.20	0.32
total								0.95	0.63	0.76

Tabela 3: Quantidade de fatos detetados pelo sistema. Os resultados apresentados para o conjunto de documentos de cada município são: a quantidade total de fatos corretamente detetados e, entre parêntesis, o número total de fatos encontrados pela pessoa, nesses mesmos documentos. Adicionalmente existem em ([1]) 4 processo de obra incorretamente extraídos e em ([2]) 1 protocolo incorretamente extraído.

não reportada).

4 Conclusões

Este artigo apresenta pela primeira vez o sistema completo com a nova interface em linguagem natural escrita. O artigo descreve ainda a criação do domínio de conhecimento e dos exemplos semente com um maior nível de detalhe em relação a publicações anteriores, permitindo assim ter-se uma percepção mais aprofundada dos procedimentos a efetuar para utilizar o sistema em casos concretos. A descrição efetuada contempla todos os módulos do sistema, proporcionando-se deste modo uma visão global do mesmo.

A aplicação desenvolvida adiciona informação semântica ao conteúdo existente em documentos escritos numa linguagem natural, o Português, e disponibiliza essa informação via uma interface em linguagem natural ou via protocolos abertos de acesso a dados. As suas principais caraterísticas são: aceitar diversos domínios do conhecimento desde que definido por uma ontologia, obter informações acerca desse domínio em textos escritos em linguagem natural, e disponibilizar a informação via interfaces apropriadas para pessoas e para máquinas.

A preparação da aplicação a um novo domínio implica um conjunto reduzido de tarefas que incluem a definição desse domínio e o fornecimento de alguns exemplos do conteúdo desse domínio nos textos a processar. É igualmente possível alterar a língua base do sistema de Português para outra alterando o módulo de PLN, sem ser necessário alterar os restantes módulos da aplicação.

Este tipo de aplicações são importantes para o e-gov porque o seu próprio sucesso depende, em grande medida, da facilidade de obtenção de informação e utilização dos seus serviços. Contudo, o desenvolvimento deste tipo de aplicações para o e-gov e para Português é uma tarefa que ainda apresenta desafios. Um deles é a adaptação a esta área específica, uma vez que o e-gov contém documentos que abarcam diversos assuntos e que, frequentemente, contêm frases de difícil interpretação devido à sua extensão e ao estilo de escrita. Outro desafio é o desenvolvimento de sistemas de extração e disponibilização de informação semântica para Português que, apesar da maturidade de vários recursos e ferramentas disponíveis, ainda não são comuns trabalhos acerca da sua integração e utilização em aplicações concretas.

Para concluir, o sistema funciona para o Português e foi construído reutilizando software do estado da arte maioritariamente desenvolvido visando o Inglês. Isto mostra que é possível - e deve ser tentado - integrar ferramentas de software de alto desempenho mesmo que inicialmente tenham sido concebidas para outras línguas naturais.

Agradecimentos

Os autores gostariam de agradecer ao Ciro Martins pela cuidada anotação dos exemplos semente nos documentos de treino do sistema.

Referências

Banko, Michele, Michael J. Cafarella, Stephen Soderland, Matthew Broadhead, e Oren Etzioni. 2007. Open information extraction from the Web. Em *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, Hyderabad, India.

Bast, Holger, Alexandru Chitea, Fabian Suchanek, e Ingmar Weber. 2007. ESTER: Efficient Search on Text, Entities, and Relations. Em *Proc. 30th ACM SIGIR*, pp. 679–686.

Berners-Lee, Tim, James Hendler, e Ora Lassila. 2001. The Semantic Web. *Scientific American*, 284(5):34–43.

Bizer, Christian, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, e Sebastian Hellmann. 2009. DBpedia - A crystallization point for the Web of Data. *Web Semantics: Science, Services and Agents on the WWW*, 7(3):154–165.

Brickley, Dan e Libby Miller. 2010. FOAF Vocabulary Specification. Publicado online em 9 de Agosto May 24th, 2010 at `http://xmlns.com/foaf/spec/`.

Cardoso, Nuno. 2008. REMBRANDT - Reconhecimento de Entidades Mencionadas Baseado em Relações e ANálise Detalhada do Texto. Em *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca.

Chakravarthy, A., F. Ciravegna, e V. Lanfranchi. 2006. Cross-media document annotation and enrichment. Em *Proc. 1st Semantic Web Authoring and Annotation Workshop (SAAW2006)*.

Chatzidimitriou, M. e A. Koumpis. 2008. Marketing One-stop e-Government Solutions: the European OneStopGov Project. *IAENG International Journal of Computer Science*, 35(1):74–79.

Chaves, M.S., M.J. Silva, e B. Martins. 2005. A Geographic Knowledge Base for Semantic Web Applications. Em *Proc. of Simpósio Brasileiro de Banco de Dados*.

Etzioni, O., M. Cafarella, D. Downey, S. Kok, A.M. Popescu, T. Shaked, S. Soderland, D.S. Weld, e A. Yates. 2004. Web-scale information extraction in KnowItAll:(preliminary results). Em *Proceedings of the 13th international conference on World Wide Web*, pp. 100–110. ACM.

Ferreira, Liliana, César Telmo Oliveira, António Teixeira, e João Paulo Silva Cunha. 2009. Extracção de Informação de Relatórios Médicos. *Linguamática*, 1(1).

Freitas, Cláudia, Paulo Rocha, e Eckhard Bick. 2008. Floresta Sintá (c) tica: Bigger, Thicker and Easier. *Computational Processing of the Portuguese Language*.

Frissen, Valerie, Jeremy Millard, Noor Huijboom, Jonas Svava Iversen, Linda Kool, Bas Kotterink, Marc van Lieshout, Mildo van Staden, e Patrick van der Duin. 2007. The Future of eGovernment: An exploration of ICT-driven models of eGovernment for the EU in 2020.

GeoNames. 2010. GeoNames Geographical Database. `http://www.geonames.org/export`.

Gruber, Thomas R. 1993. A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition*, 5:199–220.

Hall, Johan, Jens Nilsson, Joakim Nivre, Gülşen Eryiğit, Beáta Megyesi, Mattias Nilsson, e Markus Saers. 2007. Single Malt or Blended? A Study in Multilingual Parser Optimization. Em *Proc. of the Conference on Empirical Methods in Natural Language Processing and on Computational Natural Language Learning*.

Kaufmann, Esther e Abraham Bernstein. 2007. How Useful Are Natural Language Interfaces to the Semantic Web for Casual End-Users? Em Karl Aberer, Key-Sun Choi, Natasha Fridman Noy, Dean Allemang, Kyung-Il Lee, Lyndon J. B. Nixon, Jennifer Golbeck, Peter Mika, Diana Maynard, Riichiro Mizoguchi, Guus Schreiber, e Philippe Cudré-Mauroux, editores, *ISWC/ASWC*, volume 4825 of *Lecture Notes in Computer Science*, pp. 281–294. Springer.

Layne, Karen e Jungwoo Lee. 2001. Developing fully functional E-government: A four stage model. *Government Information Quarterly*, 18(2):122–136.

Li, Yunyao, Huahai Yang, e H. V. Jagadish. 2005. NaLIX: an interactive natural language interface for querying XML. Em *Proc. of the ACM SIGMOD international conference on Management of data*, pp. 902.

Mihalcea, R. 2010. Performance Analysis of a Part of Speech Tagging Task. *Computational Linguistics and Intelligent Text Processing*, pp. 299–321.

Mota, Cristina e Diana Santos, editores. 2008. *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca.

National Imagery and Mapping Agency. 2000. Department of Defense World Geodetic System 1984: its definition and relationships with local geodetic systems. `http://earth-info.nga.mil/GandG/publications/tr8350.2/tr8350_2.html`.

Oliveira, Hugo Gonçalo, Diana Santos, e Paulo Gomes. 2010. Extracção de relações semânticas entre palavras a partir de um dicionário: o PAPEL e sua avaliação. *Linguamática*, 2(1):77–93.

Paiva Dias, Gonçalo. 2006. *Arquitectura de suporte á integração de serviços no governo electrónico*. Tese de doutoramento, Universidade de Aveiro.

Ranchhod, Elisabete, Cristina Mota, e Jorge Baptista. 1999. A Computational Lexicon of Portuguese for Automatic Text Parsing. Em *Proc. of SIGLEX99: Standardizing Lexical Resources - ACL*.

Rodrigues, Mário, Gonçalo Paiva Dias, e António Teixeira. 2010a. Human Language Technologies for e-Gov. Em *Proc. of the 6th International Conference on Web Information Systems and Technologies*, pp. 400–403, Valencia, Spain.

Rodrigues, Mário, Gonçalo Paiva Dias, e António Teixeira. 2010b. Knowledge Extraction from Minutes of Portuguese Municipalities Meetings. Em *Proc. of the FALA 2010 - VI Jornadas en Tecnología del Habla and II Iberian SLTech Workshop*.

Rodrigues, Mário, Gonçalo Paiva Dias, e António Teixeira. 2011. Ontology Driven Knowledge Extraction System with Application in e-Government. Em *Proc. of the 15th Portuguese Conference on Artificial Intelligence*, pp. 760–774, Lisboa, Portugal.

Santos, Diana e Nuno Cardoso, editores. 2007. *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*. Linguateca.

Schmid, Helmut. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. Em *Proc. of International Conference on New Methods in Language Processing*, volume 12.

Sirin, E., B. Parsia, B.C. Grau, A. Kalyanpur, e Y. Katz. 2007. Pellet: A Practical OWL-DL Reasoner. *Web Semantics: science, services and agents on the World Wide Web*, 5(2):51–53.

Sroga, Magdalena. 2008. Access-eGov-Personal Assistant of Public Services. Em *Proc. of the International Multiconference on Computer Science and Information Technology*, pp. 421–427.

Suchanek, Fabian M., Gjergji Kasneci, e Gerhard Weikum. 2007. YAGO: a core of semantic knowledge. Em *WWW '07*, pp. 697–706, New York, NY, USA. ACM.

Suchanek, F.M., G. Ifrim, e G. Weikum. 2006. LEILA: Learning to Extract Information by Linguistic Analysis. Em *Proc. of the ACL Workshop OLP*.

United Nations. 2010. United Nations E-Government Survey 2010 - Leveraging e-government at a time of financial and economic crisis.

Wang, C., M. Xiong, Q. Zhou, e Y. Yu. 2007. Panto: A portable natural language interface to ontologies. *LNCS*, 4519:473.

Weibel, S., J. Kunze, C. Lagoze, e M. Wolf. 2007. Dublin Core Metadata for Resource Discovery. RFC 5013 (Informational). `http://www.ietf.org/rfc/rfc5013.txt`.

Wu, Fei, Raphael Hoffmann, e Daniel S. Weld. 2008. Information extraction from Wikipedia: moving down the long tail. Em *Proc. of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pp. 731–739, New York, NY, USA. ACM.