Capítulo 11

REMBRANDT - Reconhecimento de Entidades Mencionadas Baseado em Relações e ANálise Detalhada do Texto

Nuno Cardoso

Cristina Mota e Diana Santos, editoras, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, 2008, Capítulo 11, p. 195–211.

O REMBRANDT (**R**econhecimento de **E**ntidades **M**encionadas **B**aseado em **R**elações e **AN**álise **D**etalhada do **T**exto) é um sistema de reconhecimento de entidades mencionadas (REM) e de detecção de relações entre entidades (DRE), projectado para reconhecer todo o tipo de entidades mencionadas (EM) em textos escritos em português. O REMBRANDT explora intensamente a Wikipédia como fonte de conhecimento, e aplica um conjunto de regras gramaticais que aproveitam os vários indícios internos e externos das EM para extrair o seu significado (McDonald, 1996).

O REMBRANDT foi desenvolvido no âmbito do meu doutoramento, que foca o problema de reformulação automática de consultas realizadas a um motor de busca de âmbito geográfico, de uma forma semântica (Cardoso, 2008), e faz parte da linha de investigação seguida no projecto GReaSE (http://xldb.di.fc.ul.pt/wiki/Grease), que procura dotar os motores de busca com capacidade de raciocínio geográfico (Silva et al., 2006). O REMBRANDT nasce da necessidade de desenvolver uma ferramenta de anotação de texto capaz de reconhecer EM que possuam uma forte ligação a locais geográficos, como é o caso de nomes de países, cidades, rios, universidades, monumentos ou sedes de organizações. As aplicações do REMBRANDT envolvem a detecção de âmbitos geográficos nas consultas dos utilizadores, e a geração de "*assinaturas geográficas*" dos documentos, ou seja, listas de EM geográficas que traduzem o âmbito geográfico de cada um dos documentos, e que são usadas na recuperação e ordenação de documentos segundo critérios geográficos.

A tarefa de REM inclui vários desafios, e em relação às pretensões do REMBRANDT o problema de desambiguação de sentidos merece particular destaque; os nomes geográficos podem ser usados em diversos contextos, como é o caso de nomes de pessoas (por exemplo, *Camilo Castelo Branco*) ou de organizações (por exemplo, *France Press*), podem ser usados de forma metonímica (por exemplo, *Varsóvia* para citar o pacto) e até podem designar entidades geográficas bem diferentes (por exemplo, *Cuba* designa um país e uma cidade portuguesa). Santos e Chaves (2006) fazem uma análise sobre os contextos das EM geográficas. Assim sendo, os objectivos do REMBRANDT não se limitam ao reconhecimento de EM geográficas, abrangendo portanto todas as EM relevantes no texto precisamente para facilitar o processo de desambiguação de EM geográficas e para poder situar melhor o contexto da EM.

O REMBRANDT está disponível a todos de forma gratuita, incluindo o código fonte, sob a licença GPL em http://xldb.di.fc.ul.pt/Rembrandt.

11.1 Inspiração para o REMBRANDT

A estratégia dependente da língua do REMBRANDT foi inspirada em parte pelo sistema de REM criado por Bick (2007), o PALAVRAS_NER, e que obteve os melhores resultados nas tarefas de identificação e de classificação de EM do primeiro evento de avaliação do primeiro HAREM, em Abril de 2006.

O sistema PALAVRAS_NER baseia-se no analisador morfossintáctico PALAVRAS (Bick, 2003), que usa uma sintaxe própria para criar regras manuais que exploram os indícios das EM no texto. Contudo, as semelhanças entre o PALAVRAS_NER e o REMBRANDT acabam por aqui, uma vez que o REMBRANDT usa um sistema próprio de criação e aplicação de regras gramaticais, e utiliza a Wikipédia como base de conhecimento para a classificação de EM.

A ideia de usar a Wikipédia em vez de um almanaque para assistir o REMBRANDT na sua tarefa surge através dos trabalhos recentes em torno da Wikipédia, e que evidenciam

as potencialidades que este recurso possui para as tarefas de extracção de informação (Wu e Weld, 2007). Um exemplo disso é o trabalho de Auer e Lehmann (2007), que explora as caixas de informação (em inglês, *infoboxes*) das páginas da Wikipédia para gerar conhecimento em forma de factos representados por triplas em RDF. O projecto associado ao seu trabalho, o `DBpedia.org`, contava em 2008 com cerca de 100 milhões de triplas RDF extraídas automaticamente a partir da Wikipédia (Auer et al., 2007).

O REMBRANDT foi inicialmente concebido para usar a Wikipédia como se se tratasse de um simples almanaque, mais actualizado e vasto do que os almanaques usados por outros sistemas de REM. No entanto, o funcionamento do REMBRANDT rapidamente evoluiu para tirar partido da riqueza da informação e estrutura da Wikipédia, permitindo inclusive a prospecção de informação adicional sobre cada EM, como acontece no caso de extracção de informação geográfica implícita (Cardoso et al., 2008b).

O REMBRANDT possui uma interface própria para interagir com a Wikipédia, a SASKIA, com o objectivo de facilitar as tarefas de navegação na estrutura de categorias, ligações e redireccionamentos da Wikipédia com vista à extracção de conhecimento. Já existe, por exemplo, o RENOIR (Santos et al., 2008a), que é uma ferramenta de construção de consultas semânticas que usa a API da SASKIA para executar consultas específicas à colecção da Wikipédia.

11.2 Anatomia do REMBRANDT

O REMBRANDT suporta vários formatos de ficheiros, na leitura e na escrita (texto simples, HTML ou XML), e pode ser executado em qualquer plataforma que possua uma máquina virtual de Java. Para processamento de grandes quantidades de texto, o REMBRANDT pode funcionar em regime de mapeamento e redução (em inglês, *MapReduce*) (Dean e Ghemawat, 2008) através da sua extensão para o Apache Hadoop (`http://hadoop.apache.org`), permitindo a distribuição das tarefas de REM por vários computadores disponíveis.

A figura 11.1 resume o funcionamento do REMBRANDT. Os documentos são tratados, um de cada vez, numa linha de processos de anotação sucessivos até à sua versão final e definitiva, tal como por exemplo em Gruhl et al. (2004). Ao longo da linha de processos, as EM entretanto reconhecidas vão guardando um historial de alterações desde que são detectadas pela primeira vez até à sua última modificação. Este sistema de rastreio de EM facilita a depuração das acções de cada processo, permitindo a afinação do sistema de regras e leis a aplicar a cada EM, bem como vigiar as suas aplicações ao longo da vida de cada EM. O funcionamento do REMBRANDT pode ser subdividido em três etapas principais:

i. Reconhecimento de expressões numéricas e geração de candidatas a EM

Os textos são previamente divididos em frases e em unidades, com a ajuda do atomizador da Linguateca (disponível através do módulo de Perl `Lingua::PT::PLN`, em `http://search.cpan.org/~ambs/Lingua-PT-PLN-0.17`). Um primeiro conjunto de regras identifica expressões numéricas no texto, tais como unidades compostas só por algarismos, ou números por extenso, ordinais e cardinais. De seguida são aplicadas regras para reconhecer expressões temporais e valores, tirando proveito dos números já reconhecidos no passo anterior.

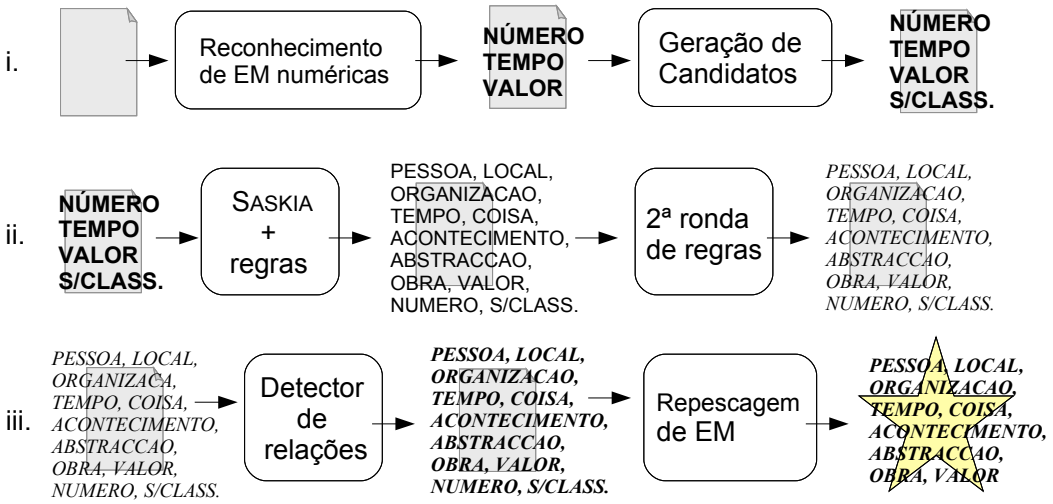

Figura 11.1: O funcionamento do REMBRANDT, dividido em três etapas.

A geração de candidatas a EM é feita através da identificação de sequências de unidades contendo pelo menos uma letra maiúscula e/ou um algarismo, podendo existir uma das seguintes unidades, desde que não comece ou termine a EM: *de*, *da*, *do*, *das*, *dos*, *e* (doravante designados por unidades *daeose*, devido à sua expressão regular `'d[aeo]s?|e'`).

ii. Classificação de EM

Na segunda etapa, cada uma das candidatas a EM é classificada primeiro pela SASKIA, e depois é novamente classificada através de regras gramaticais. Esta estratégia de "dupla classificação" das EM tem um conjunto de vantagens: primeiro, a SASKIA realiza uma classificação de acordo com vários significados que a EM pode ter, e que são reunidos nas páginas típicas de desambiguação da Wikipédia. Desta forma, cria-se um ponto de partida a partir do qual o processo de desambiguação pode trabalhar com vista à selecção do significado correcto da EM. Segundo, as regras gramaticais englobam indícios externos e internos das EM, o que de certa forma supervisiona as classificações da SASKIA segundo o contexto da EM.

Para ilustrar o funcionamento das regras, considere o exemplo dado pela frase (11.1), onde a SASKIA classificou previamente a EM *Angola* como sendo `LOCAL/HUMANO/PAIS`. A aplicação de uma regra gramatical dedicada à captura de ruas vai mudar a classificação da EM *Angola* para `LOCAL/HUMANO/RUA`, devido à presença da expressão *Rua da* antes da EM.

(11.1) Eu moro na Rua de Angola.

A terminar a etapa de classificação, é aplicada uma segunda ronda de regras gramaticais, que aproveita as classificações existentes para detectar EM com uma morfologia mais elaborada. É nesta fase que, por exemplo, as EM que possuem um termo *daeose* são analisadas na sua eligibilidade para serem representadas através de uma etiqueta `<ALT>`,

ou se porventura é melhor serem divididas em EM mais pequenas, que serão novamente classificadas pela SASKIA e pelas regras gramaticais.

iii. Repescagem de EM sem classificação

Na última etapa, realiza-se a detecção de relações entre EM através de um conjunto de regras específicas para a tarefa. As relações entretando detectadas são usadas para repescar algumas EM sem classificação, mas que estão relacionadas com EM devidamente classificadas.

Após a tarefa de DRE, é feita uma última repescagem de EM com nomes de pessoas, através de uma comparação com uma lista de nomes comuns. Por último, as EM que persistem sem classificação são eliminadas, bem como números por extenso sem uma letra maiúscula (uma vez que não são considerados EM segundo as directivas em vigor, excepto no caso de expressões temporais). As EM de categoria `NUMERO` são convertidas em `VALOR/QUANTIDADE`.

11.3 SASKIA

A SASKIA é a interface responsável por pré-processar as colecções da Wikipédia, e por realizar uma classificação inicial às EM, com base na informação extraída da Wikipédia. A API da SASKIA permite realizar operações simples de interacção com a colecção da Wikipédia, como por exemplo a navegação nas páginas, a extracção de categorias, a recolha e filtragem de âncoras ou a normalização de títulos das páginas.

11.3.1 Pré-processamento da Wikipédia

A Wikipédia gera periodicamente imagens estáticas dos conteúdos relativos a cada língua, disponibilizadas em `http://download.wikipedia.org`, onde podem ser acedidas pelo público em geral. Estas imagens são compostas por vários ficheiros em XML e ficheiros em SQL, consoante o nível de informação associada (por exemplo, a inclusão ou não das páginas de discussão, das páginas dos utilizadpres, ou do histórico de alterações das páginas).

Os ficheiros em XML contêm o texto das páginas no seu formato MediaWiki original (`http://meta.wikimedia.org/wiki/Help:Editing`), enquanto os ficheiros em SQL incluem o código para a criação das tabelas (metadados das páginas, ligações entre páginas, informação das categorias e tabela de redireccionamentos) e os respectivos dados.

A SASKIA foi desenvolvida inicialmente em torno do ficheiro em XML das páginas portuguesas da imagem estática da Wikipédia em português. O pré-processamento do ficheiro era feito através de uma versão modificada do programa Wikipedia Preprocessor (`http://sourceforge.net/projects/wikiprep`), extraindo as ligações, o texto das âncoras, as categorias, os títulos, os subtítulos, as listas ordenadas, as páginas relacionadas, os URL externos e o texto filtrado de cada documento. A informação extraída era armazenada e indexada através do Lucene (`http://lucene.apache.org`).

No entanto, esta estratégia serve perfeitamente para imagens da Wikipédia semelhantes à portuguesa, cujo ficheiro em XML (de cerca de 1,4 GB de tamanho, em Março de 2008) é pré-processado em poucas horas; para imagens como a versão inglesa da Wikipédia, com cerca de 28 GB de tamanho, em Fevereiro de 2008, o tempo de pré-processamento é proibitivamente longo. Assim sendo, após a participação no segundo HAREM, a SASKIA

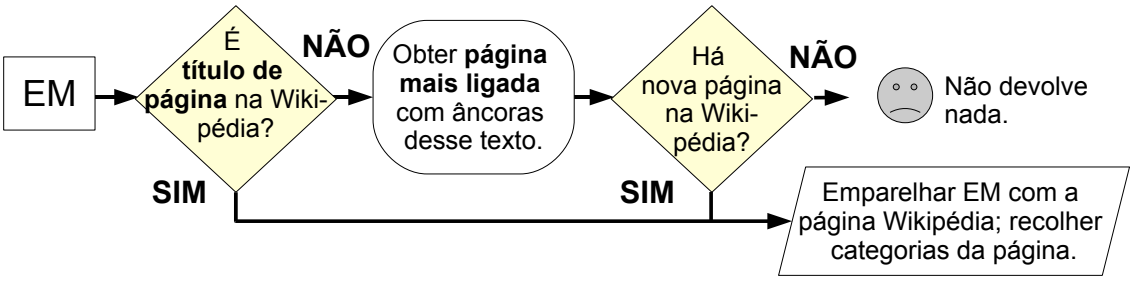

Figura 11.2: Emparelhamento de EM realizado pela SASKIA.

foi melhorada de forma a explorar as imagens da Wikipédia em formato SQL, podendo ser pré-configurada de forma a escolher a fonte de informação (XML ou SQL) a usar para cada uma das acções da sua API (a partir da versão 0.8 do REMBRANDT).

11.3.2 Estratégia de classificação

O procedimento de classificação da SASKIA usado no Segundo HAREM está dividido em três etapas: (i) associação da EM a uma página da Wikipédia, (ii) recolha de categorias associadas à EM, e (iii) mapeamento das categorias da Wikipédia às classificações do HAREM. Para evitar consultas repetidas à SASKIA, esta possui uma memória temporária (em inglês, *cache*) interna que guarda os resultados das classificações anteriores, acelerando significativamente o tempo de resposta para EM já analisadas previamente.

i. Emparelhamento de EM

A SASKIA começa por procurar uma página da Wikipédia com o título exactamente igual ao texto da EM. Se for encontrada, passa-se para a etapa seguinte; se não for encontrada, o texto da EM é usado para encontrar a página mais ligada através de âncoras cujo texto é idêntico ao texto da EM (ver figura 11.2).

O uso das ligações entre páginas da Wikipédia permite à SASKIA lidar com as várias formas de designação de uma mesma entidade. Um exemplo é as diversas formas de designar a entidade *Estados Unidos da América* (país): *EUA*, *USA*, *Estados Unidos*, *E.U.A.* ou *América do Norte*, que são tudo formas vulgares e abreviadas de referir a mesma entidade. Uma vez que a grande maioria das ligações da Wikipédia com o texto da âncora contendo estas variantes aponta para a página (`http://pt.wikipedia.org/wiki/Estados_Unidos_da_América`), o emparelhamento das EM faz-se de uma forma simples e robusta.

Esta estratégia de análise do texto das âncoras só pode ser usada se a SASKIA tiver ao seu dispor a versão pré-processada em XML da Wikipédia. Caso se opte por usar a versão em SQL da Wikipédia para realizar o emparelhamento das EM, a SASKIA recorre à tabela de redireccionamentos para encontrar a página respectiva. A principal desvantagem desta opção é que o emparelhamento fica dependente da existência de uma entrada de redireccionamento explícito na tabela SQL.

ii. Recolha de categorias

Para cada uma das categorias da página da Wikipédia emparelhada na etapa anterior, a SASKIA analisa o seu tipo e, caso seja necessário, visita mais páginas relacionadas e extrai as suas categorias, adicionando-as à lista. A SASKIA adopta uma estratégia de profundidade-primeiro na sua navegação entre páginas, limitada até ao quarto nível de profundidade (ver figura 11.3). Os tipos de categoria da Wikipédia que a SASKIA reconhece são os seguintes:

Categoria normal. Estas categorias são simplesmente adicionadas à lista de categorias.

Autocategoria. Designo por autocategoria toda a categoria que possui o mesmo nome do título da página que a contém. Por exemplo, a página da Wikipédia da cidade do Porto (`http://pt.wikipedia.org/wiki/Porto`) possui a autocategoria `Categoria:Porto`. Nestes casos, a SASKIA analisa a página da categoria (`http://pt.wikipedia.org/wiki/Categoria:Porto`) e adiciona as suas categorias à lista.

Categoria de desambiguação. A `Categoria:Desambiguação` é usada nas páginas da Wikipédia que esclarecem os diversos significados da EM, e que reúnem ligações para as respectivas páginas desambiguadas. Nestes casos, a SASKIA extrai as ligações da página que possuam o texto da EM na âncora (com base no XML), ou as ligações para páginas da Wikipédia cujo título contém o texto da EM (com base no SQL), visita as páginas referenciadas e recolhe as suas categorias.

Categoria de acrónimo. A `Categoria:Acrónimos` é usada nas páginas da Wikipédia cujo título é um acrónimo. A SASKIA procura a presença desta categoria nas páginas de desambiguação, para que a extracção de ligações para outras páginas se faça sem usar o texto da EM (que é um acrónimo). Exceptuando este caso, esta categoria não é utilizada para nenhum fim, e não é adicionada à lista de categorias.

Um exemplo ilustrativo de uma página com as categorias `Acrónimos` e `Desambiguação` é a página da Wikipédia da PSP (`http://pt.wikipedia.org/wiki/PSP`). Sendo uma página de desambiguação, as suas ligações apontam para páginas sobre entidades com significados distintos, como a Polícia de Segurança Pública, a PSP PlayStation Portable ou o Paint Shop Pro.

Contudo, como a sigla *PSP* é um acrónimo, o texto da EM não pode ser directamente usado para filtrar as ligações. Nesses casos, a SASKIA compara o acrónimo e a ligação, e determina se o texto da âncora (com base no XML) ou o título da página-alvo (com base no SQL) representa uma expansão do acrónimo. Se tal se verificar, a nova página é então visitada e as suas categorias são recolhidas.

A utilização do texto das EM para filtrar as ligações da página de desambiguação é essencial, uma vez que nem todas as suas ligações são relevantes (por exemplo, pode haver uma ligação para a página de Portugal na frase de descrição sumária do sentido da Polícia de Segurança Pública).

iii. Classificação das categorias

Finalmente, a SASKIA aplica uma lista de regras gramaticais específicas sobre cada uma das categorias, com o objectivo de extrair o seu significado e a sua referência geográfica,

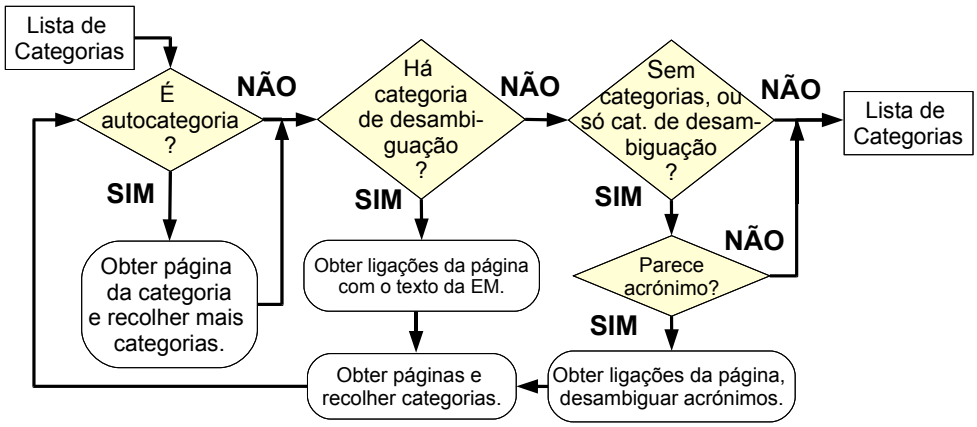

Figura 11.3: Recolha de categorias pela SASKIA.

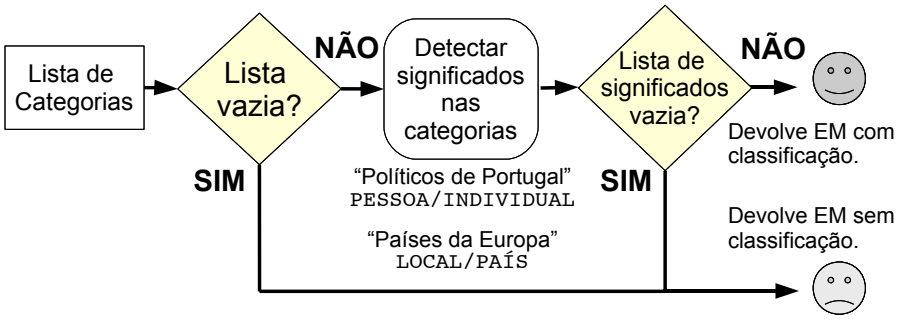

Figura 11.4: Classificação das categorias pela SASKIA.

caso exista (ver figura 11.4). Por exemplo, a categoria *Cantores de Portugal* possui uma morfologia frequentemente usada na Wikipédia portuguesa, para o qual foi criada uma regra que procura uma definição, seguida de um termo *daeose* opcional e de uma entidade geográfica (tanto como nome, como na sua variante em adjectivo, como acontece com a categoria *antigas Províncias portuguesas*).

Após a aplicação válida desta regra, a definição *cantores* é mapeada à respectiva classificação do HAREM (`PESSOA/INDIVIDUAL`) com a ajuda de um almanaque de definições interno. A entidade geográfica *Portugal* é recolhida, desambiguada e associada à EM como se tratasse de informação geográfica implícita sobre o âmbito geográfico da EM, conforme descrito por Cardoso et al. (2008b).

11.4 Regras gramaticais

As regras gramaticais representam padrões nas frases que indiciam a presença de EM com determinadas propriedades semânticas, e definem as acções a tomar quando estas são apli-

cadas com sucesso. As regras são compostas por uma ou mais cláusulas, ou seja, unidades de padrões mais simples.

As cláusulas são aplicadas ordenadamente, uma de cada vez, a uma parte seleccionada da frase. Cada cláusula retorna verdade se as unidades alinhadas com esta corresponderem ao seu padrão, ou retorna falso no caso oposto. Se todas as cláusulas retornarem verdade, a regra diz-se bem sucedida e retorna por sua vez um valor verdadeiro. Em contraste, se pelo menos uma das cláusulas retornar falso, ou se a frase terminar e ainda houver cláusulas obrigatórias para aplicar, a regra falha e retorna um valor falso.

Quando a regra é bem sucedida e retorna verdade, segue-se a execução de uma acção pré-determinada, definida pelo campo **Acção**. Desta forma, é possível definir regras com actuações diferentes, como é o caso de regras de DRE, regras de geração de `<ALT>` ou regras de geração de novas EM.

As regras gramaticais usadas na etapa de classificação de EM (regras de detecção de indícios internos e externos) possuem o campo **Acção:GerarEM** para que a sua aplicação resulte em novas EM com novas classificações definidas na regra através dos campos **categoria**, **tipo** e **subtipo**. Adicionalmente, o campo **PolíticaDaRegra** define a política de escolha das unidades que irão fazer parte da nova regra, e pode tomar dois valores: i) **Regra**, o que inclui todas as unidades capturados por todas as cláusulas da regra (normalmente usado para regras de índicio interno) e ii) **Cláusula**, onde cada cláusula especifica se as unidades capturadas por ela vão ser incluídas ou não na nova EM (normalmente usado para regras de indício externo).

11.4.1 Propriedades das cláusulas

As cláusulas também possuem propriedades próprias que descrevem a sua forma de actuação. As propriedades mais importantes de uma cláusula são as seguintes:

Cardinalidade, que define se a cláusula é obrigatória ou opcional, e determina o número de vezes que pode ser aplicada. A cardinalidade pode tomar os seguintes valores: i) **Zero ou um**, semelhante à semântica de '.?' das expressões regulares. A cláusula é opcional, e devolve verdadeiro quer tenha sido correspondida ou não a um conjunto de unidades. ii) **Zero ou mais**, semelhante à semântica de '.*?' das expressões regulares. A cláusula é opcional, e devolve verdadeiro independentemente das vezes que conseguir ser correspondida. iii) **Um**, semelhante à semântica de '.' das expressões regulares. A cláusula é obrigatória e é executada uma única vez, devolvendo verdadeiro se for correspondida. iv) **Um ou mais**, semelhante à semântica '.+' das expressões regulares, onde é obrigatório haver pelo menos uma correspondência verdadeira. As cláusulas adoptam uma estratégia gananciosa (em inglês, *greedy*), procurando a maior sequência de unidades possível, antes de passar para a cláusula seguinte (se existir).

Critério, que define o tipo de correspondência, e pode tomar os seguintes valores: i) **Simples**, que faz uma comparação simples entre unidades, ii) **Expressão**, que aplica uma expressão regular a um termo, iii) **EM**, que procura a presença de uma EM com um determinado leque de classificações, e iv) **Conceito**, que usa uma lista de conceitos, comparando cada elemento dessa lista, um por um, até encontrar uma expressão que corresponda às unidades.

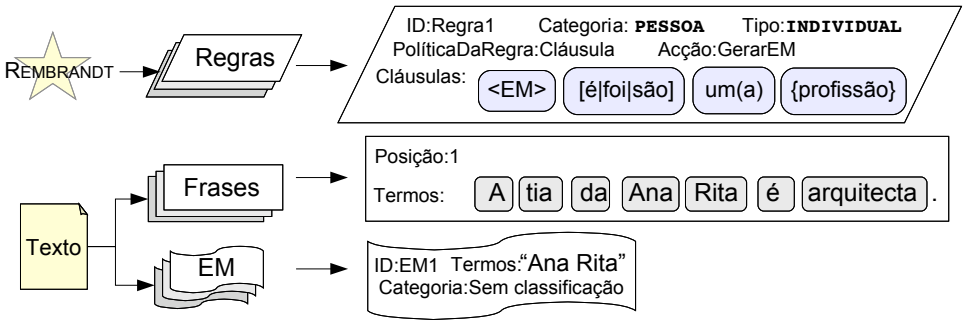

Figura 11.5: Selecção de regras gramaticais, frases e de EM.

Padrão, que instancia o padrão a ser aplicado na comparação, de acordo com o critério. Assim sendo, o padrão pode incluir um termo, uma expressão regular, uma lista de classificações, ou uma lista de conceitos (ou seja, uma lista de listas de expressões regulares).

Inclusão, que define se as unidades correspondidas pela cláusula irão fazer parte da EM, no caso da regra ter o campo **Acção:GerarEM**. Este campo é lido só se a respectiva regra definir o campo **PolíticaDaRegra:Cláusula**.

11.4.2 Aplicação das regras

A aplicação das regras ao texto é feita de uma forma sequencial, uma regra de cada vez, a todas as frases do texto (também de forma sequencial, da primeira para a última frase). Para cada frase do texto, a regra activa começa pelo primeiro termo da frase, e invoca sucessivamente cada uma das cláusulas. Após esse passo, a regra muda o seu posicionamento para um termo à direita, até serem esgotadas todas as combinações possíveis de alinhamento da regra com a frase. As regras bem sucedidas são logo executadas, e no caso das regras de geração de EM, as novas EM ficam imediatamente disponíveis para serem usadas na aplicação das regras seguintes.

Esta forma de encadeamento de regras de acordo com a sua ordem inicial permite a elaboração de regras sequenciais. Por exemplo, para capturar a EM *entre Abril e Maio*, é usada uma primeira regra que reconhece os meses, e depois é aplicada uma segunda regra, que procura um padrão `entre <EM> e <EM>` para então juntar as unidades todas numa nova EM.

A figura 11.5 ilustra a aplicação da regra gramatical com o identificador `Regra_1` à frase (11.2). Note-se que a EM *Ana Rita*, previamente reconhecida como candidata a EM e que se encontra de momento sem classificação, foi invocada para esta aplicação pois faz parte da frase. As propriedades da `Regra_1` determinam que, caso seja bem sucedida, irá gerar uma nova EM que terá a classificação de `PESSOA/INDIVIDUAL`, e que a escolha das unidades da nova EM será feita pelas cláusulas.

(11.2) A tia da Ana Rita é arquitecta.

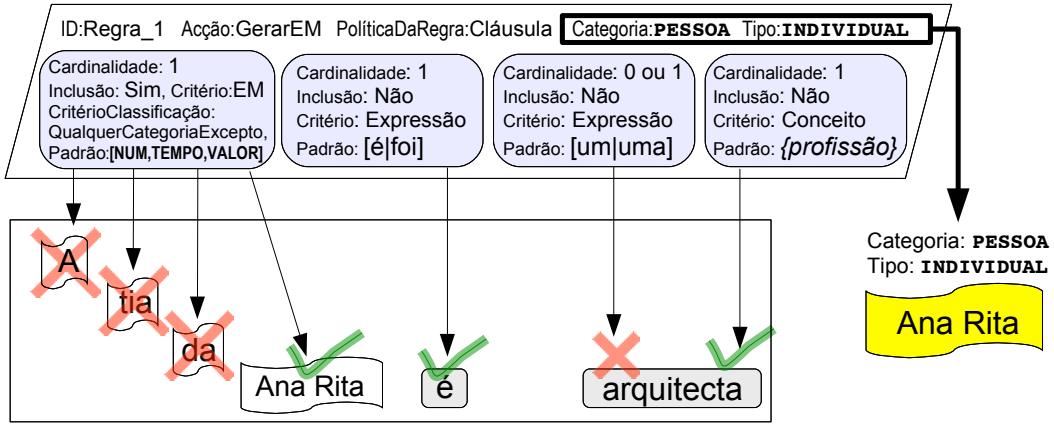

Figura 11.6: Aplicação de regras gramaticais.

A aplicação da regra começa com o alinhamento da primeira cláusula com o início da frase. Esta cláusula procura encontrar uma EM (**Critério:EM**) que não tenha nenhuma das seguintes classificações: `NÚMERO`, `TEMPO` e `VALOR` (**CritérioClassificação:QualquerCategoriaExcepto**). Como tal, a cláusula irá falhar sucessivamente quando alinhada às unidades *A*, *tia* e *da*, uma vez que não fazem parte de nenhuma EM (sempre que uma cláusula falha, a regra é então repetida com um novo alinhamento à direita, geralmente de um termo). Finalmente, a EM *Ana Rita* é então correspondida pela primeira cláusula, e os seus duas unidades são guardadas devido ao campo **Inclusão:Sim**.

De seguida, a regra passa para a cláusula seguinte, que é alinhada ao termo seguinte. Esta segunda cláusula procura um padrão no termo que corresponda a *é* ou *foi*, e é obrigatório encontrar esse termo para que a regra seja bem sucedida. Como o termo *é* é encontrado, é a vez da terceira cláusula, que é opcional visto que possui o campo **Cardinalidade:Zero ou Um**. Este tipo de cláusulas são úteis para representar pequenas variações nas morfologias das frases que se procura detectar, como é o caso de *(...) é uma arquitecta* ou *(...) é arquitecta*. As cláusulas opcionais retornam sempre positivas, independentemente de conseguirem ser correspondidas às unidades.

Finalmente, a quarta e última cláusula, de **Critério:Conceito**, procura a definição de uma profissão/ocupação através de uma lista de expressões regulares que podem ser simples (por exemplo, `['[Aa]rquitec?t[oa]s?']`) ou compostas (por exemplo, `['[Tt][éê]cnic[oa]s?','[Oo]ficia[li]s?', 'de', '[Cc]ontas']`). Quando uma das expressões regulares é correspondida, a regra verifica que não há mais cláusulas a satisfazer e retorna com sucesso, gerando então a nova EM `<EM CATEG=¨PESSOA¨ TIPO=¨INDIVIDUAL¨>Ana Rita</EM>` (ver figura 11.6).

11.4.3 Tribunal de EM

Quando as regras geram novas EM que se sobrepõem a EM já existentes, diz-se que há um "*conflito*" entre EM. Assim sendo, o REMBRANDT possui um tribunal de EM, onde os

conflitos entre EM são resolvidos. No tribunal, a EM "*ré*", que já existia na lista de EM do documento, e a EM "*acusadora*", recém-gerada pela regra, esgrimem argumentos para que se possa decidir o seu destino. O tribunal dispõe de um conjunto de leis de resolução de conflito, de onde é seleccionada a lei mais adequada para o conflito em questão, e do qual sai um veredicto que pode incluir desde a eliminação de uma das EM, da geração de alternativas `<ALT>`, até à fusão das EM numa única.

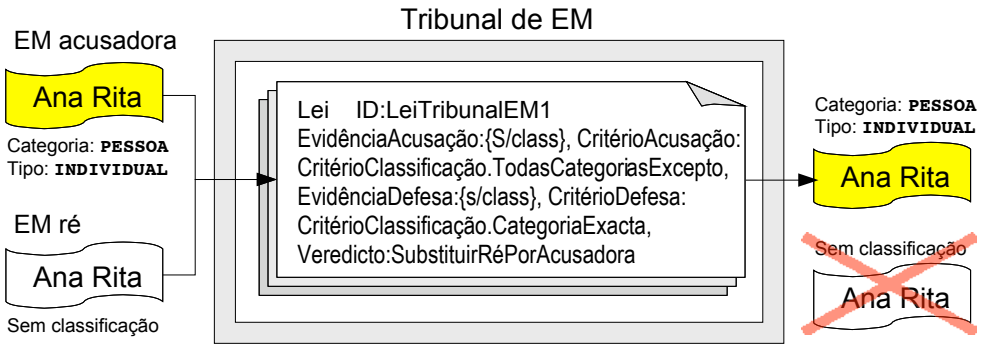

Figura 11.7: Exemplo de aplicações de uma lei no tribunal de EM.

A figura 11.7 ilustra a aplicação das leis para a situação retratada na figura 11.6. A lei aplicada diz que uma EM sem classificação deverá sempre ser substituída por uma EM equivalente com uma classificação válida. Os indícios das EM consideradas nas leis incluem leques de classificações e formas de sobreposição entre EM em conflito (ou seja, se uma EM está contida noutra, se sobreposta, se está contida e ajustada à esquerda, entre outros).

O tribunal permite uma certa organização e prioritização das EM geradas pelas regras. No entanto, a passagem pelo tribunal é um comportamento por omissão, feito se não houver indicação em contrário na regra; se for necessário, é possível definir o campo **PolíticaConflito** na regra, para definir previamente o veredicto a tomar em caso de conflitos. Este campo serve, por exemplo, para aplicar em regras de captura de `<ALT>`, onde não há propriamente um conflito entre EM, mas sim uma interpretação alternativa do sentido das EM envolvidas.

11.5 Detecção de relações entre EM

O sistema de detecção de relações do REMBRANDT usa heurísticas básicas de relacionamento entre EM com base nas suas unidades, nas suas categorias e nas ligações das respectivas páginas da Wikipédia. As heurísticas são aplicadas a EM não-numéricas (isto é, sem classificação `VALOR`, `NUMERO` ou `TEMPO`) e seguem o seguinte procedimento:

1. EM com o mesmo texto são rotuladas como sendo idênticas (`ident`). As EM que foram emparelhadas à mesma página da Wikipédia também são rotuladas como sendo idênticas; desta forma, EM como por exemplo *Cavaco Silva* e *Aníbal Cavaco Silva* são associadas com a etiqueta `ident`.

2. EM que se sobrepõem a outras EM (no caso de `<ALT>`) ou que são separadas por um termo *daeose* são analisadas nas suas classificações, que determinam o tipo de relação entre elas. Por exemplo, a relação `ocorre_em` é usada quando uma EM com categoria `ACONTECIMENTO` se sobrepõe ou é vizinha de uma EM de categoria `LOCAL`, como acontece em *Jogos Olímpicos de Pequim*; a relação `sede_em` é usada na mesma situação, mas com uma EM com categoria `CONSTRUCAO`, como acontece em *Museu Militar do Porto*. No final são repescadas relações `ident` a EM que possuem texto em comum e alinhado a um extremo, como acontece com nomes de pessoas, por exemplo, *José Sócrates* e *Sócrates*.

3. EM que estejam emparelhadas a páginas da Wikipédia são analisadas de forma a encontrar relacionamentos com EM vizinhas na mesma frase, através das ligações da página. Os textos das âncoras da página (usando a Wikipédia em XML) ou os títulos das páginas-alvo (usando a Wikipédia em SQL) podem indiciar uma relação entre as EM, como é ilustrado pelo exemplo das EM *Neil Armstrong* e *NASA*; uma vez que a página da Wikipédia do astronauta contém uma ligação para a página da NASA, é adicionada uma relação `outra` entre estas duas EM.

4. Finalmente, é aplicada uma série de regras gramaticais vocacionadas para detectar relações entre EM numa mesma frase e que ainda não possuem relações entre elas. Essas regras gramaticais definem o campo **Acção:GerarRelação**, e possuem cláusulas com o campo **Papel**, que identificam o papel de cada uma das EM visadas, e o tipo de relacionamento detectado.

O mecanismo de detecção de relações entre EM do REMBRANDT ainda está nos seus passos iniciais, no entanto o seu papel será determinante para a desambiguação de sentidos de EM com várias classificações. Por exemplo, a EM *Armstrong* é classificada pela SASKIA como sendo um local e uma pessoa ao mesmo tempo; contudo, se na sua vizinhança existir a EM *NASA*, a detecção de relações pode assinalar que afinal o seu sentido é de um nome de pessoa.

11.6 Resultados no Segundo HAREM

O REMBRANDT enviou um total de três corridas para o Segundo HAREM. A fonte de informação usada pela SASKIA foi o ficheiro em XML relativo à imagem estática da Wikipédia de 2 de Março de 2008, que conta com 405.752 páginas e 5.010.715 ligações. A geração de corridas foi realizada no regime de mapeamento e redução do Hadoop v0.15, num grupo (em inglês, *cluster*) de 7 máquinas Linux com 19 processos de mapeamento e 7 processos de redução, tendo demorado uma média de 100 minutos para etiquetar a colecção HAREM.

Na altura do Segundo HAREM, a etapa de DRE foi a mais pesada em termos de processamento do REMBRANDT chegando a contribuir com mais de metade do tempo de processamento. Para este facto contribui a falta de optimização do sistema de DRE, que procura relações para todas as combinações de pares de EM possíveis que ainda não possuem uma relação, fazendo com que o tempo de operação evolua de forma exponencial em relação ao número de EM do documento, e chegando a vários minutos num documento longo.

11.6.1 Corridas

As três corridas foram geradas por diferentes versões 0.7 do REMBRANDT. As diferenças entre versões limitaram-se à rectificação de alguns erros no funcionamento do REMBRANDT e ao melhoramento da SASKIA e das regras gramaticais, após uma análise das saídas geradas. Em resumo:

11.6.1.0.1 Corrida REMBRANDT_1. Gerada pelo REMBRANDT 0.7.1, possui uma versão não testada de regras gramaticais focadas na detecção de <ALT>, uma nova interface de escrita de <ALT>, e o sistema de detecção de relações parcialmente programado, mas ainda não afinado quanto à sua estratégia adoptada e nas regras gramaticais usadas.

11.6.1.0.2 Corrida REMBRANDT_2. Gerada pelo REMBRANDT 0.7.2, possui melhoramentos da SASKIA a nível da estratégia em páginas de desambiguação e de acrónimos, e foram afinadas as regras próprias para os <ALT>, abrangendo mais casos elegíveis. O sistema de relações foi aperfeiçoado de maneira a propagar a eliminação de relações, o que acontecia frequentemente devido à eliminação de EM sem classificação na última etapa, deixando órfãs muitas das relações encontradas.

11.6.1.0.3 Corrida REMBRANDT_3. Gerada pelo REMBRANDT 0.7.3, possui um sistema de detecção de relações afinado com um leque final de regras. Foram rectificados vários problemas a nível da escrita da saída. Ao nível da SASKIA, a navegação entre páginas para a recolha de categorias realiza-se agora com profundidade quatro em vez de três, e possui melhorias a nível de resolução de acrónimos.

11.6.2 Resultados na tarefa de REM

Corrida	Avaliação estrita de ALT			Avaliação relaxada de ALT		
	Precisão	Abrangência	Medida F	Precisão	Abrangência	Medida F
REMBRANDT_2	0,6497	0,5036	0,5674	0,6622	0,5173	0,5808
REMBRANDT_3	0,6286	0,5032	0,5590	0,6424	0,5163	0,5725
REMBRANDT_1	0,6396	0,4690	0,5412	0,6505	0,4809	0,5530

Tabela 11.1: Resultados do REMBRANDT no HAREM clássico, no cenário total.

A tabela 11.1 apresenta os resultados globais do REMBRANDT, no HAREM clássico. A corrida REMBRANDT_2 obteve os melhores valores, o que é um resultado curioso, uma vez que a corrida REMBRANDT_3 foi gerada com o propósito de rectificar vários problemas observados na corrida REMBRANDT_2.

A tabela 11.2 apresenta os resultados do REMBRANDT discriminados por categoria. Salienta-se o facto de o REMBRANDT ter tido um bom desempenho na categoria LOCAL e no respectivo cenário selectivo que incluía somente EM de classificação LOCAL/HUMANO e LOCAL/FISICO, o que é particularmente relevante do ponto de vista da sua aplicação em sistemas de recuperação de informação geográfica.

Os resultados para as EM de categoria PESSOA beneficiaram bastante dos passos de repescagem de EM, quer pela detecção de relações, quer pelo uso de um almanaque de

Categoria	Melhor corrida	Classificação			Identificação		
		P	A	F	P	A	F
PESSOA	REMBRANDT_3	0,7683	0,5368	0,6320	0,7747	0,5411	0.6371
LOCAL	REMBRANDT_1, 3	0,5484	0,6607	0,5993	0,5553	0,7241	0,6286
VALOR	REMBRANDT_3	0,4127	0,7176	0,5241	0,4161	0,7247	0,5287
TEMPO	REMBRANDT_3	0,5904	0,4030	0,4790	0,6098	0,4093	0,4899
ORGANIZACAO	REMBRANDT_3	0,5350	0,3231	0,4029	0,6035	0,3624	0,4529
OBRA	REMBRANDT_3	0,5251	0,2171	0,3072	0,5276	0,2188	0,3093
ACONTECIMENTO	REMBRANDT_3	0,5630	0,2026	0,2980	0,6312	0,2242	0,3308
ABSTRACCAO	REMBRANDT_3	0,1956	0,1433	0,1655	0,2085	0,1534	0,1768
COISA	REMBRANDT_2	0,0451	0,0566	0,0502	0,0425	0,0724	0,0536

Tabela 11.2: Resultados do REMBRANDT discriminados por categoria, e ordenados pela medida F da tarefa de classificação.

nomes, uma vez que a SASKIA está mais vocacionada para reconhecer nomes de celebridades, e as regras gramaticais revelaram-se insuficientes para abranger os variados indícios externos de nomes de pessoas.

No caso oposto, os resultados para as EM de categoria VALOR saíram algo prejudicados pela conversão automática de EM de categoria NUMERO para VALOR/QUANTIDADE. Este problema também se reflecte na abrangência das EM de categoria TEMPO, muito por culpa da dificuldade em detectar o verdadeiro significado dos números que representam anos. O quinteto de categorias com melhores resultados é fechado pela categoria ORGANIZACAO, que não teve o desempenho que se esperava devido à estratégia algo simplista de geração de candidatas a EM, que precisa de ser revista e adaptar-se para outras morfologias de EM.

11.6.3 Resultados na tarefa de DRE

Os resultados da pista do ReRelEM são analisados de uma forma global no capítulo 4. A tabela 11.3 apresenta os resultados obtidos pelo REMBRANDT na tarefa de detecção de relações entre EM, o ReRelEM. O cenário total diz respeito a todas as EM presentes na CD, enquanto que o cenário selectivo 5 diz respeito às EM de categoria LOCAL e de tipo HUMANO ou FISICO, ou seja, EM de cariz geográfico.

Cenário	Medida	Melhor corrida	Total			Melhor corrida	Selectivo 5		
			P	A	F		P	A	F
Todas	Relações	REMB. 1	0,5822	0,3669	0,4502	REMB. 1	0,9178	0,6204	0,7403
Identidade	Relações	REMB. 1	0,7723	0,6934	0,7307	REMB. 3	0,9184	0,9000	0,9091
Inclusão	Relações	REMB. 1	0,3236	0,3261	0,3243	REMB. 1	0,9615	0,4098	0,5747
Localização	Relações	REMB. 2	0,4048	0,1288	0,1954	REMB. 1	-	-	-

Tabela 11.3: Resultados do REMBRANDT na pista do ReRelEM.

No geral, a corrida REMBRANDT_1 obteve os melhores resultados em DRE se considerarmos a medida F. As outras duas corridas, apesar de obterem valores de medida F próximos, nota-se que sacrificaram a precisão para aumentar de forma ténue a abrangência, o que in-

dica que as alterações introduzidas nas corridas REMBRANDT_2 e REMBRANDT_3 também introduziram muito ruído.

Em mais detalhe, nota-se que o REMBRANDT é eficaz na detecção de relações de identidade para todo o tipo de EM (cerca de 0,73 de medida F), mas não na detecção de relações de inclusão (0,32) e de localização (0,19). Contudo, para as EM de cariz geográfico os valores são mais elevados, com cerca de 0,9 na identificação de relações entre entidades geográficas, e 0,57 na detecção de inclusões.

Em resumo, o REMBRANDT destaca-se pela positiva na tarefa de detecção de relações para entidades geográficas, o que é encorajador dado os propósitos do REMBRANDT de extracção de pistas geográficas do texto. O desempenho do REMBRANDT neste capítulo ainda possui uma margem de progressão considerável, face ao conjunto simples de regras de DRE usadas.

11.7 Conclusões e trabalho futuro

O REMBRANDT é um sistema de REM muito ambicioso, propondo-se etiquetar todo o tipo de EM existentes no texto, e detectar o tipo de relacionamento entre elas, a partir de uma estratégia de regras gramaticais manuais co-adjuvadas por um sistema de extracção de conhecimento automático a partir da Wikipédia, a SASKIA. Assim sendo, é essencial acompanhar a evolução do seu desempenho ao longo do seu desenvolvimento, de forma a afinar adequadamente todos os diversos passos que compõem a linha de processamento do REMBRANDT.

A participação do REMBRANDT no Segundo HAREM reveste-se de particular importância, pois permitiu ter uma primeira noção do desempenho do REMBRANDT nas suas tarefas, em particular do seu nível de eficiência em relação ao reconhecimento de entidades geográficas, e à detecção de relações entre elas. Os resultados obtidos são animadores e mostram que a estratégia adoptada pelo REMBRANDT permite obter desempenhos satisfatórios em REM.

Após a participação no Segundo HAREM, o REMBRANDT já foi melhorado em diversos aspectos, e há uma lista de melhoramentos a realizar no REMBRANDT a curto e médio prazo, nomeadamente:

Abstracção da camada de classificação. O REMBRANDT foi desenvolvido em torno da hierarquia de classificação adoptadas pelo Segundo HAREM, estando esta codificada de raiz no funcionamento do REMBRANDT. Para trabalho futuro, o REMBRANDT irá suportar diferentes leques de categorização de EM, permitindo a sua adaptação a domínios mais específicos e o aumento da resolução da classificação.

Adaptação para várias línguas. O REMBRANDT v0.8 já suporta várias línguas na sua tarefa de REM, embora não possua ainda forma de detectar automaticamente a língua do documento. A adaptação a outras línguas requer a re-escrita das regras e das leis do REMBRANDT e a readaptação das definições da Wikipédia às características da imagem respectiva. Actualmente, o REMBRANDT já processa textos na língua inglesa, embora o seu desempenho esteja ainda aquém do desempenho observado para o português.

Detecção de contextos. O REMBRANDT precisa de uma "3ª ronda" de regras gramaticais específicas para capturar o contexto mais genérico das EM, de forma a adaptar-se

melhor à metodologia de REM sugerida pelo HAREM. Um exemplo típico é a utilização de entidades geográficas em contextos abstractos, como é patente na frase *A honra da França estava em jogo*, que o HAREM reconhece como sendo uma ABSTRACCAO/IDEIA. Uma vez que a EM *França* não é referida num papel geográfico, tal facto tem de transparecer na forma de actuação do REMBRANDT. Em estudo está a possibilidade de esta 3ª ronda de regras ser realizada através de métodos de aprendizagem automática, uma vez que é difícil representar o contexto de forma objectiva e contundente através de regras gramaticais.

Pré-processamento da Wikipédia. A SASKIA precisa de optimizar o seu processo de pré-processamento dos ficheiros de XML, de forma a lidar convenientemente com ficheiros de grandes dimensões. Para tal, está prevista a implementação de um pré-processador próprio, em vez de usar uma ferramenta externa, que consiga capturar e organizar a informação contida nas caixas de informação. A SASKIA também está a ser melhorada de forma a usar uma camada abstracta de representação dos documentos, de forma a que o seu funcionamento não dependa do tipo de ficheiro (XML ou SQL) usado no pré-processamento, e permitindo a exploração de outras fontes de informação (a DBpedia, por exemplo).

Melhorias na API da SASKIA. A SASKIA deverá ser capaz de explorar mais informação a partir das páginas da Wikipédia, como é o caso de coordenadas geográficas, os primeiros parágrafos do texto, ou as caixas de informação. Adicionalmente, a SASKIA terá de se adaptar às novas tendências de organização de categorias da Wikipédia, onde começa a ser comum encontrar categorias que são constituídas por subcategorias, numa hierarquia de dois níveis que é algo semelhante à categorização usada no HAREM.

Re-utilização de conhecimento adquirido durante a anotação. O âmbito de acção do REMBRANDT está restringido ao nível do documento, ou seja, não há transposição de conhecimento adquirido entre documentos. O REMBRANDT poderá tirar partido de um centro de armazenamento de conhecimento, onde poderá guardar informação importante sobre EM normalmente usadas, e desta forma agilizar a anotação de novos documentos. Por exemplo, a EM HAREM poderá estar explicitamente descrita num determinado documento, mas noutro documento o seu significado poderá ser muito difícil de extrair, devido à falta de indícios no texto.

Desambiguação de sentidos a partir da DRE. A fraca abrangência obtida pelo REMBRANDT na tarefa ReRelEM indicia que existe uma margem de progresso considerável nesta fase. O REMBRANDT irá depender em grande parte da detecção de relações para a desambiguação de sentidos das EM, em particular das EM geográficas.