Reconhecimento automático de relações entre entidades mencionadas em textos de língua portuguesa

Mírian Bruckschen[1,2], Renata Vieira[2], Sandro Rigo[1,3]

[1]Universidade do Vale do Rio dos Sinos (UNISINOS)

[2]Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS)

[3]Universidade Federal do Rio Grande do Sul (UFRGS)

mirian.bruckschen@gmail.com, renata.vieira@pucrs.br, rigo@unisinos.br

Resumo. *Este trabalho apresenta um sistema para o reconhecimento automático de relações semânticas entre entidades mencionadas (entidades com nomes próprios) em textos de língua portuguesa.*

Abstract. *This paper presents a tool for automatic recognition of semantic relations between named entities (entities with proper names) within texts in portuguese language.*

Palavras-chave: processamento de linguagem natural; entidades mencionadas; relações semânticas; língua portuguesa

1. Introdução

O reconhecimento de entidades mencionadas e relações entre estas é uma importante questão para a Lingüística Computacional. De acordo com a definição que utilizamos neste trabalho, entidades mencionadas são aquelas referenciadas no texto por um nome próprio. Considera-se que a tarefa de identificação e classificação destas entidades é um primeiro e importante passo na análise semântica de textos (Santos e Cardoso, 2007).

Numa iniciativa pioneira no cenário de processamento da língua portuguesa, é proposto o HAREM[1] (Linguateca, 2008), uma avaliação conjunta de sistemas reconhecedores de EMs[2]. Na edição atual, foi criada uma trilha de reconhecimento de relações entre EMs (ReRelEM[3]), na qual participamos com o presente trabalho, um sistema que reconhece relações entre EMs identificadas pelo analisador PALAVRAS (Bick, 2000).

Neste trabalho, é contextualizado o problema do reconhecimento automático de relações entre entidades no texto, e a ferramenta participante na trilha ReRelEM do HAREM, SeRELeP[4], é apresentada. Os resultados oficiais da avaliação conjunta são discutidos, bem como possibilidades de aplicações do sistema.

---

[1]HAREM é uma Avaliação de Reconhecedores de Entidades Mencionadas, conforme disponível em: <http://www.linguateca.pt/HAREM/>

[2]Entidades Mencionadas

[3]Reconhecimento de Relações entre EMs, conforme disponível em <http://acdc.linguateca.pt/aval_conjunta/HAREM/ReRelEM.html>

O restante deste artigo está distribuído da seguinte forma: a Seção 2 lista e discute brevemente trabalhos e áreas relacionadas; a Seção 3 apresenta a ferramenta desenvolvida para o reconhecimento de relações, ferramentas auxiliares utilizadas e todo o processo de anotação da coleção disponibilizada pelo HAREM com EMS e suas relações, assim como os resultados obtidos; a Seção 4 delineia aplicações práticas para este trabalho; e, finalmente, a Seção 5 encerra o trabalho com considerações finais e trabalhos futuros.

2. Conceitos e trabalhos relacionados

É possível observar um grande número de esforços no sentido de criar ferramentas, técnicas e outros recursos para o processamento da linguagem. Em algumas línguas, como o inglês, a disponibilidade destes recursos é maior. No caso da língua portuguesa, no entanto, ainda está em andamento a construção e disponibilização destes recursos.

Além disso, a avaliação de resultados em português costumava ser um processo difícil e subjetivo, dado que não existiam bases de dados de texto ou técnicas comuns para comparação. Na língua inglesa, atividades de avaliação conjunta já são realizadas há algum tempo. Alguns exemplos são o MUC[5] e, mais recentemente, o ACE[6]. O ACE ainda ocorre, e em sua edição atual propõe, dentre outras tarefas, o reconhecimento de entidades mencionadas e relações semânticas entre estas, à semelhança da tarefa do HAREM que motivou o presente trabalho. (NIST e ACE, 2007).

O HAREM é uma avaliação conjunta do processo de reconhecimento de entidades mencionadas (*named entities*, que são as entidades referenciadas por nome próprio). O HAREM está em sua segunda edição (Linguateca, 2008), e a importância da avaliação de técnicas e ferramentas de forma padronizada é evidente para o avanço da área de processamento da linguagem (Santos e Cardoso, 2007).

No que se refere à identificação automática da relação de identidade entre entidades em um texto (resolução de correferência), este é outro tópico dentro da Linguística Computacional que tem tido especial atenção de diversos especialistas tanto das áreas de Linguística como de Informática. Isso é devido principalmente à sua aplicabilidade nas importantes e relativamente recentes áreas de extração de informação, sistemas de respostas e construção automática de ontologias, relacionadas à *Web* Semântica (Vieira et al. 2000; Soon et al, 2001; Souza, 2007).

3. SeRELeP: Sistema de reconhecimento de RELações em textos de Língua Portuguesa

Esta seção descreve o o modelo semântico do Segundo HAREM e o funcionamento do SeRELeP, a ferramenta desenvolvida neste trabalho. Além disso, são apresentados os resultados oficiais obtidos pela ferramenta na avaliação.

---

[4]Sistema de reconhecimento de RELações em textos de Língua Portuguesa

[5]*Message Understanding Conference*, conforme disponível em <http://www-nlpir.nist.gov/related_projects/muc/>

[6]*Automatic Content Extraction*, conforme disponível em <http://www.nist.gov/speech/tests/ace/>

3.1. Modelo do HAREM e o Segundo HAREM

O HAREM é uma avaliação conjunta com tarefas bastante abrangentes. A tarefa principal trata-se do reconhecimento de entidades mencionadas no texto, e a tarefa extra que é contemplada neste trabalho é a identificação de relações entre estas entidades. Estas duas tarefas serão descritas, em conjunto com o modelo do HAREM, nesta subseção.

A proposta do HAREM é que todas as entidades mencionadas sejam identificadas e classificadas. Sua classificação não é obrigatória, mas encorajada. As categorias de EMs no Segundo HAREM são: PESSOA, ORGANIZAÇÃO, ACONTECIMENTO (que chamamos EVENTO), LOCAL, ABSTRAÇÃO, OBRA, VALOR, COISA, OUTRO. A categoria OUTRO é uma forma de contemplar entidades que não se encaixem nas demais, para que não seja necessário deixar de classificar alguma entidade.

Outras avaliações conjuntas, como o antigo MUC, partem duma visão diferente da seguida pelo HAREM. Assumindo alguns objetos de interesse (pessoas, organizações e lugares, por exemplo), propõe-se o mapeamento de entidades deste tipo, descartando outras. A Figura 1, de (Santos e Cardoso, 2007), ilustra esta diferença de forma bastante clara.

No HAREM, as EMs devem ser classificadas de acordo com o sentido que adquirem no texto, isto é, considerando seu contexto. Assim como Brasil é um país (LOCAL), pode também ser uma seleção de futebol (grupo da classe PESSOA) ou uma idéia, como quando falamos em “um Brasil mais unido” (ABSTRAÇÃO).

O sistema SeRELeP não faz a classificação explícita das EMs em sua anotação, muito embora utilize esta informação para inferência das relações.

As relações propostas pela trilha ReRelEM, pioneira no Segundo HAREM, são: *ident* (identidade), *inclui*, *ocorre_em* e *outra*. A relação *ident* é atribuída a EMs que refiram-se ao mesmo objeto no mundo ("Carmen Miranda", "Carmen Miranda" e "Carmen", em diferentes posições do texto). Para que haja a atribuição desta relação, é necessário que as EMs pertençam à mesma categoria semântica. Se “Carmen Miranda” se referir a uma música em homenagem à cantora, por exemplo, seria da classe OBRA, e portanto não estabeleceria relação de identidade com “Carmen Miranda”, a cantora e atriz também chamada de “Pequena Notável”.

Já a segunda relação proposta, *inclui*, refere-se a EMs das classes LOCAL, ORGANIZAÇÃO, ABSTRAÇÃO, TEMPO e COISA. Estas podem ter relação de inclusão entre si (uma EM da classe ORGANIZAÇÃO pode incluir apenas outra da mesma classe, por exemplo). Um exemplo seria a inclusão de um lugar em outro, como “Brasil” *inclui* “Rio Grande do Sul”, que *inclui* “Porto Alegre”.

A relação *ocorre_em*, por sua vez, anuncia a ocorrência ou sede de EMs de EVENTO/ORGANIZAÇÃO em EMs de LOCAL. “Exército Zapatista de Libertação Nacional” *ocorre_em* “México”, e “São Leopoldo Fest” *ocorre_em* “São Leopoldo” são alguns exemplos.

Finalmente, a relação *outra* seria atribuída a todas as EMs que possuíssem alguma relação diferente das três anteriores (como em "José *outra* Pedro", sendo José pai de Pedro, por exemplo).

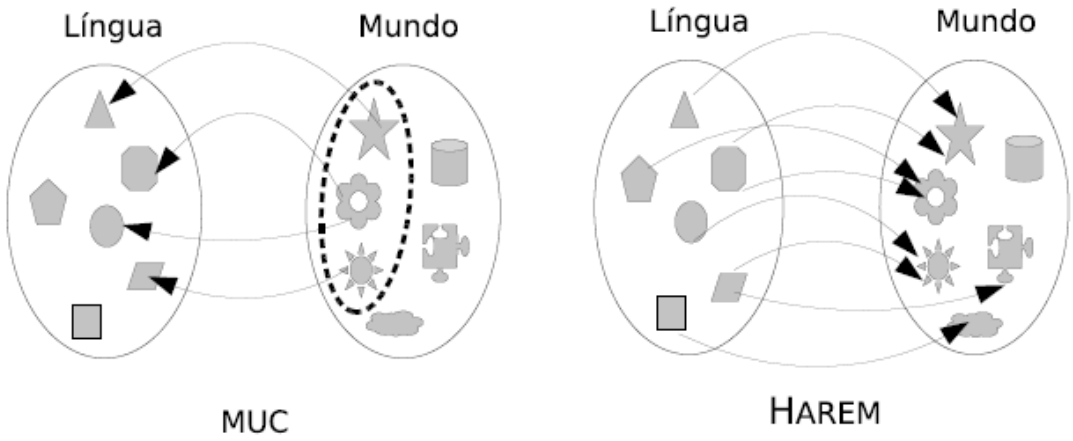

Figura 1. Pontos de partida na abordagem da semântica (Santos e Cardoso, 2007)

3.2. Visão geral do sistema

O sistema tem como entrada o arquivo de texto da coleção do HAREM em formato XML[7] e seus respectivos arquivos em formato XCES[8]. Para obtenção dos arquivos neste formato, é necessário o pré-processamento do texto pelo analisador PALAVRAS e pelo conversor Tiger2XCES (Souza, 2007b).

O PALAVRAS tem como saída, dentre outras informações, os nomes próprios contidos no texto e sua classificação semântica, que possui um mapeamento direto para a classificação no modelo do HAREM (a ser detalhado na próxima subseção). O Exemplo 1 ilustra um trecho de texto com EMs reconhecidas pelo PALAVRAS.

São Leopoldo é uma cidade localizada na região metropolitana de **Porto Alegre**, no **Rio Grande do Sul**. O estado é conhecido por suas tradições, belas modelos e desenvolvimento econômico. Dentre as diversas atrações da cidade, localiza-se em **SL** o **Museu do Trem**, um museu ferroviário, e um teatro recém inaugurado pela prefeitura junto à biblioteca municipal. Anualmente ocorre também na cidade a **São Leopoldo Fest**, festa que se dá em homenagem à chegada dos imigrantes alemães fundadores da cidade, e que reúne participantes de todo o estado. O atual prefeito da cidade, **Ary Vanazzi**, tem sido reconhecido pela população como atuante e preocupado também com a qualidade de vida nos bairros mais pobres, para os quais não existia acesso fácil à cultura e lazer.

Exemplo 1. Texto de exemplo com EMs reconhecidas pelo PALAVRAS

[7] eXtensible Markup Language

[8] XML CES (Corpus Encoding Standard for XML), conforme disponível em http://www.xces.org

O conversor Tiger2XCES converte os arquivos retornados pelo PALAVRAS, em formato TigerXML, em arquivos no formato XCES. Essa conversão ocorre sem perda de informação. A única transformação, além do formato, é a divisão da informação linguística em três diferentes arquivos: *token*, *POS* (*parts of speech*), e *phrase*.

Os arquivos de *phrase* trazem informação sintática acerca do texto e delimitação deste (sentenças, sintagmas, verbos, apostos e afins). Os arquivos de *tokens*, por sua vez, trazem informação morfossintática e uma delimitação mais granular, na qual se baseiam os arquivos de *phrase*: as palavras e entidades separadas por espaços ("a", "cidade", "de", "São_Leopoldo"). Finalmente, os arquivos de *POS* trazem informações que chamamos de traços morfológicos (gênero e número de substantivos, por exemplo), além da etiquetação semântica de entidades. O Exemplo 2 ilustra estes arquivos.

```
- <cesAna version="1.0.4">
  - <struct to="1" type="token" from="0">
      <feat name="id" value="t1"/>
      <feat name="base" value="O"/>
    </struct>
  - <struct to="9" type="token" from="2">
      <feat name="id" value="t2"/>
      <feat name="base" value="técnico"/>
    </struct>
  - <struct to="18" type="token" from="10">
      <feat name="id" value="t3"/>
      <feat name="base" value="espanhol"/>
    </struct>
  - <struct to="31" type="token" from="19">
      <feat name="id" value="t4"/>
      <feat name="base" value="Jordi_Ribera"/>
    </struct>
  - <struct to="32" type="token" from="31">
      <feat name="id" value="t5"/>
      <feat name="base" value=","/>
    </struct>
```

```
- <cesAna version="1.0.4">
  - <struct type="pos">
      <feat name="id" value="pos1"/>
      <feat name="class" value="art"/>
      <feat name="tokenref" value="t1"/>
      <feat name="canon" value="o"/>
      <feat name="gender" value="M"/>
      <feat name="number" value="S"/>
    </struct>
  - <struct type="pos">
      <feat name="id" value="pos2"/>
      <feat name="class" value="n"/>
      <feat name="tokenref" value="t2"/>
      <feat name="canon" value="técnico"/>
      <feat name="gender" value="M"/>
      <feat name="number" value="S"/>
      <feat name="semantic" value="Hprof"/>
    </struct>
```

```
- <cesAna version="1.0.4">
  - <struct to="t26" type="phrase" from="t1">
      <feat name="id" value="phr1"/>
      <feat name="cat" value="s"/>
      <feat name="function" value=""/>
    </struct>
  - <struct to="t26" type="phrase" from="t1">
      <feat name="id" value="phr2"/>
      <feat name="cat" value="fcl"/>
      <feat name="function" value="STA"/>
    </struct>
  - <struct to="t4" type="phrase" from="t1">
      <feat name="id" value="phr3"/>
      <feat name="cat" value="np"/>
      <feat name="function" value="S"/>
      <feat name="head" value="t2"/>
    </struct>
```

Exemplo 2. Arquivos no formato XCES: *token*, *POS* e *phrase*

Esta etiquetação fornecida pelo PALAVRAS associa aos nomes contidos no texto alguma das categorias pré-definidas disponíveis no dicionário interno ao analisador. A partes do corpo, por exemplo, é atribuída a etiqueta "An", de "Anatomia". A profissionais (técnicos, jogadores, professores), é atribuída a etiqueta "Hprof", de "profissão Humana".

Existe um mapeamento direto da classificação atribuída pelo PALAVRAS à classificação de acordo com o Modelo do HAREM. Cada etiqueta semântica relaciona-se com uma classe de entidade no HAREM. Este mapeamento é detalhado no *website* Floresta Sinta(c)tica[9], mas citamos aqui alguns exemplos: "hum", "official" e "member", etiquetas que referem-se, respectivamente, a entidades humanas (pessoas), oficiais (pessoas com cargos públicos, por exemplo) e participantes de algum grupo maior de pessoas (como membros de clubes e organizações). Todos estes referem-se à categoria PESSOA no HAREM.

Quanto à utilização de arquivos no formato XCES ao invés do TigerXML que é a saída do PALAVRAS, esta é uma decisão de projeto. O formato TigerXML é difícil e

[9] http://beta.visl.sdu.dk/visl/pt/info/portsymbol.html

trabalhoso de analisar, ao contrário do formato XCES. O formato XCES divide as informações de acordo com seus níveis, sintático e morfológico, de forma simples e fácil para analisadores automáticos. Da mesma forma que arquivos em formato texto plano são mais simples para a leitura por humanos, arquivos em formato XCES são mais simples para o desenvolvimento de sistemas automáticos de análise morfossintática.

Finalmente, a saída do sistema SeRELeP é um arquivo com o texto já marcado com as EMs e suas relações, também em formato XML. O Exemplo 3 traz um trecho de um arquivo de saída como exemplo.

```
− <DOC DOCID="aa40383">
    A hora da verdade 1. Hoje à noite o
    <EM ID="aa40383-EM_1">Benfica</EM>
    tem um teste elucidativo sobre o seu verdadeiro valor. Teve-o também antes contra o
    <EM ID="aa40383-EM_2">Manchester United</EM>
    , onde passou com distinção, é facto, mas o
    <EM ID="aa40383-EM_3">Manchester</EM>
    desta época e, sobretudo, o de Dezembro passado, é uma pálida imitação dos «red devils» de um passado recente, que tudo
    atemorizavam. O
    <EM ID="aa40383-EM_4" COREL="aa40383-EM_35" TIPOREL="ocorre_em">Liverpool</EM>
    , para além de campeão europeu em título, é melhor equipa e mais calculista. Contra ele, o
    <EM ID="aa40383-EM_5" COREL="aa40383-EM_1 aa40383-EM_35" TIPOREL="ident ocorre_em">Benfica</EM>
    tem de jogar o seu máximo e então se verá se o seu máximo está ou não ao mais alto nível europeu. Ao iniciar com uma
    derrota inapelável este terrível ciclo de nove dias em que tudo se pode sonhar e tudo se pode perder, o
    <EM ID="aa40383-EM_6">Benfica de Ronald Koeman</EM>
    parece ter lançado a descrença entre os seus --- adeptos e jornalistas. De repente, a tão louvada equipa de há três semanas
    atrás parece já não inspirar confiança a ninguém e o seu treinador virou um saco de pancada ao alcance de todas as
    frustrações. Acusa-se
    <EM ID="aa40383-EM_7">Koeman</EM>
    de mudar constantemente a equipa, como se ele fosse responsável pelas lesões e castigos ou pela chegada de uma
    mini-enxurrada de reforços em Dezembro, anunciados pela Direcção como o suplemento que faltava para a conquista da
    <EM ID="aa40383-EM_8" COREL="aa40383-EM_35" TIPOREL="ocorre_em">Liga dos Campeões</EM>
    e que ele, logicamente, tinha de experimentar e utilizar. E esquecendo-se até de que foi uma dessas inesperadas «revoluções»
    na formação da equipa que conduziu à também inesperada vitória contra o
    <EM ID="aa40383-EM_9" COREL="aa40383-EM_3" TIPOREL="ident">Manchester</EM>
```

Exemplo 3. Arquivo de saída do SeRELeP

A Figura 2 ilustra todo o processo de anotação automática de EMs e relações da coleção do HAREM. O PALAVRAS e o Tiger2XCES são ferramentas externas necessárias ao processo. Todas as demais partes foram desenvolvidas no decorrer do trabalho apresentado neste documento.

SeRELeP é o sistema identificador de relações entre as EMs, e o SeRELeP Tools é um conjunto de pequenos programas auxiliares, necessários à etapa de pré-processamento (para o PALAVRAS e o SeRELeP). A entrada do processo é um arquivo XML no formato do HAREM, fornecido no início da participação da avaliação conjunta. Este arquivo contém diversos textos individuais.

Os textos são extraídos pelo SeRELeP Tools em dois formatos: texto plano, que é a entrada para o PALAVRAS, e XML do HAREM, que é entrada para o SeRELeP efetivamente. Além do XML do HAREM, SeRELeP ainda precisa de outra entrada, que são os textos no formato XCES já mencionados.

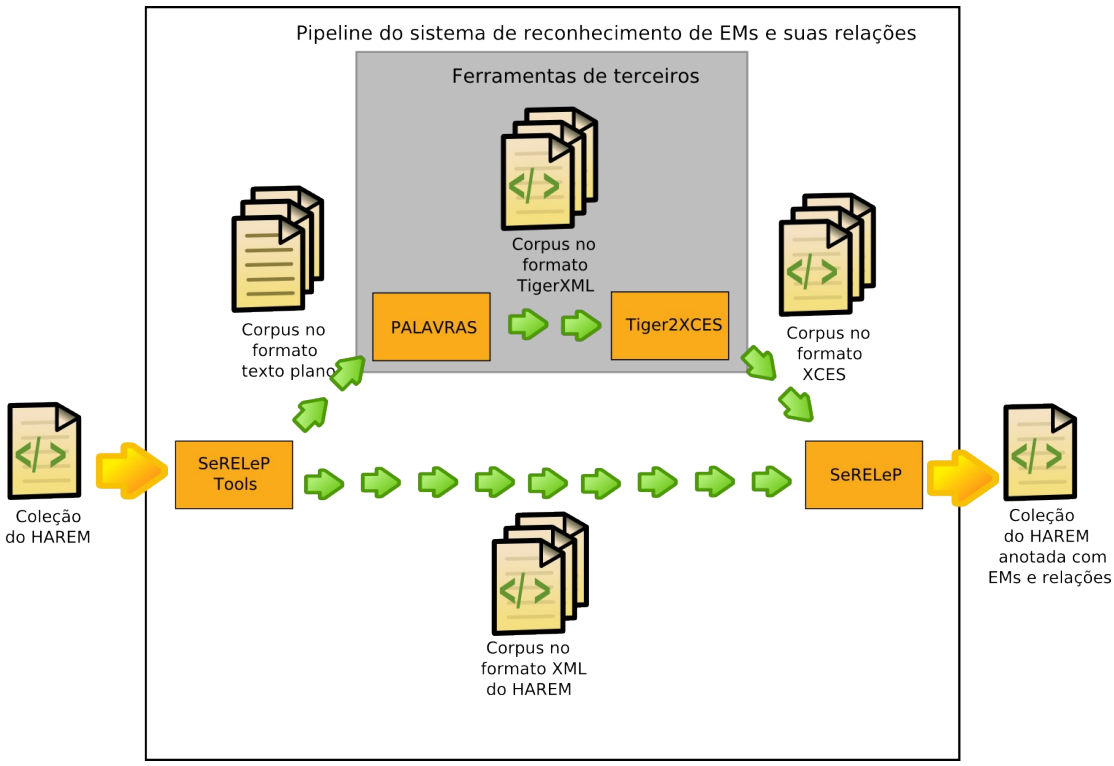

Figura 2. Processo de anotação automática de EMs e suas relações

3.3. Reconhecimento de relações

Conforme já comentado, o HAREM propõe quatro relações entre EMs: *ident* (identidade), *inclui*, *ocorre_em* e *outra*. Destas, SeRELeP trata as três primeiras. O tratamento de cada uma destas relações atualmente é detalhada mais adiante nesta seção.

Convém notar que as heurísticas descritas a seguir fazem uso da etiquetação semântica do pré-processamento realizado pelo PALAVRAS no que refere-se à categorização das EMs. Esta categorização divide as EMs nas classes semânticas propostas pelo HAREM. A Figura 3 expressa na forma de um grafo dirigido as relações *ident*, *inclui* e *ocorre_em* entre as EMs pertencentes a estas classes, conforme tratado por SeRELeP.

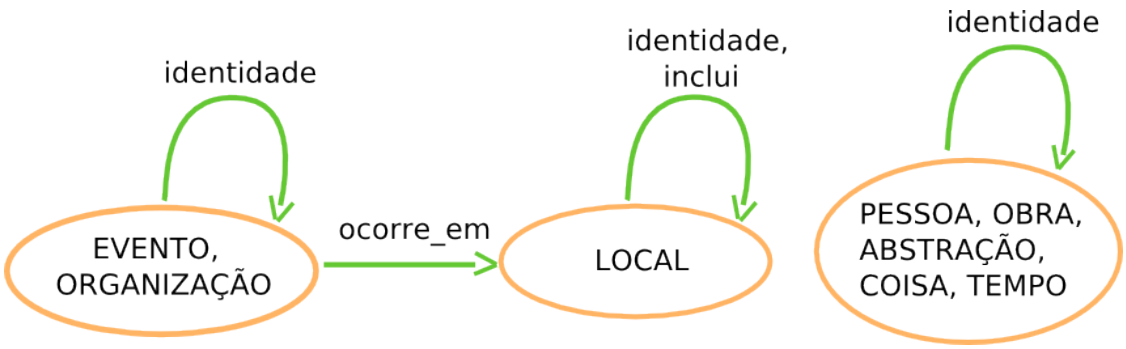

Figura 3. Relações entre as classes semânticas das EMs

A relação *ident* é atribuída a EMs que refiram-se a uma mesma entidade no mundo. A atribuição desta relação dá-se através das seguintes regras: i) comparação da EM de forma literal, isto é, se elas possuem exatamente o mesmo nome; ii) se uma é sigla da outra, isto é, se uma das EMs retoma as iniciais da outra; iii) se as EMs comparadas forem da classe PESSOA e parte do sintagma de uma for igual ao sintagma da outra (como "Carmem" e "Carmem Miranda", por exemplo). Além disso, as EMs devem pertencer à mesma categoria semântica (uma EM de EVENTO só pode estabelecer relação de identidade com outra EM de EVENTO, por exemplo).

Já a relação *inclui*, tratada somente entre EMs de LOCAL, é estabelecida mediante regras bastante simples: i) as duas EMs não podem ter relação *ident* entre si; ii) devem estar na mesma sentença; iii) deve haver uma preposição que denote inclusão (ou alguma contração de preposição que dê este significado mais artigo), como "em", "no" e "na".

Finalmente, a relação *ocorre_em* é tratada entre EMs de EVENTO e LOCAL ou de ORGANIZAÇÃO e LOCAL. As regras obedecidas por ela são verificadas na seguinte ordem: i) se houver uma EM de LOCAL cujo sintagma seja parte do sintagma da EM de EVENTO ou ORGANIZAÇÃO verificada, esta EM de EVENTO ou ORGANIZAÇÃO pode ter a si atribuída a relação *ocorre_em* referindo-se à EM de LOCAL em questão (como em "Brigada Militar de Porto Alegre *ocorre_em* Porto Alegre"); ii) se isso não acontecer, é verificada a existência de uma EM de LOCAL na mesma sentença da EM de EVENTO/ORGANIZAÇÃO analisada. Se existir, esta EM de EVENTO/ORGANIZAÇÃO será relacionada a esta EM de LOCAL através da relação *ocorre_em*; iii) se não, busca a EM de LOCAL mais próxima dentro do texto (se houver) para relacionar com a EM de EVENTO/ORGANIZAÇÃO analisada.

3.4. Resultados e discussão

Foram divulgados recentemente os resultados oficiais, tanto do HAREM clássico, que refere-se somente à identificação e classificação de EMs, quanto do ReRelEM, que refere-se à identificação e classificação das relações entre EMs. Estes resultados são descritos detalhadamente nesta seção, acompanhados de uma discussão e idéias para solucionar os problemas já identificados.

A parte de identificação e classificação de EMs é feita totalmente pelo PALAVRAS. Quanto à identificação, os resultados são mostrados na Tabela 1, e são resultados já muito interessantes, fruto da maturidade atingida pela ferramenta ao longo dos últimos anos. O PALAVRAS mantém-se entre os melhores sistemas que participaram do Segundo HAREM, na tarefa de identificação de entidades. Convém salientar que o PALAVRAS participa do HAREM desde a sua primeira edição, e certamente esta experiência conta para a obtenção de resultados tão positivos. Na coleção final anotada pelo SeRELeP, não classificamos as EMs explicitamente, somente as suas relações, e por este motivo os resultados da classificação de EMs (HAREM clássico) não são aqui exibidos.

	Precisão	**Abrangência**	**Medida F**
Identificação	82%	60%	69%

Tabela 1. Resultados oficiais do HAREM clássico (PALAVRAS)

A anotação realizada pelo PALAVRAS é utilizada pelo sistema SeRELeP na tarefa de reconhecimento de relações. Os resultados desta tarefa são ilustrados na Tabela 2, que traz tanto a identificação da relação como sua classificação (em *ident*, *inclui*, *ocorre_em* ou *outra*). A Tabela 3 ilustra os resultados individuais por relação tratada.

	Precisão	Abrangência	Medida F
Identificação	68%	35%	46%
Classificação	57%	30%	39%

Tabela 2. Resultados oficiais da trilha ReRelEM (SeRELeP)

	Precisão	Abrangência	Medida F
ident	87%	54%	66%
inclui	56%	11%	19%
ocorre_em	28%	21%	24%

Tabela 3. Resultados oficiais da trilha ReRelEM por relação (SeRELeP)

A relação com melhores resultados é claramente a *ident*. Atribui-se este desempenho ao fato de que regras simples, como as utilizadas e descritas neste documento, já abrangem boa parte das relações entre EMs.

Algumas questões ainda devem ser tratadas em maior detalhe, como apelidos não relacionados ao nome original (um exemplo seria "Pequena Notável" e "Carmem Miranda", que têm relação de identidade não-detectada pelo sistema) e o mesmo nome com pequenas diferenças de grafia, comumente causados por erros de digitação ("Maria de Costa" e "Maria da Costa"). Quanto às demais relações, *inclui* e *ocorre_em*, os resultados são inferiores.

Um dos problemas que atingiu o desempenho do sistema em todas as relações é a categorização semântica efetuada no pré-processamento. Apesar da tarefa de delimitação de EMs ter sido muito bem-sucedida, a classificação não o foi, e isto afetou diretamente o reconhecimento das relações.

Um exemplo disso é a marcação de EMs de LOCAL tais como "Biblioteca Victor Civita" como ORGANIZAÇÃO. Em "Carmen Miranda conquistou a Broadway", a EM "Broadway" não é indicada na coleção dourada como um lugar, mas sim um grupo de pessoas, pertencendo à categoria semântica PESSOA, e não LOCAL, já de acordo com o analisador sintático, a classificação resultante é LOCAL.

Pretende-se tratar algumas destas questões inicialmente com a utilização de três técnicas: i) busca e inferência em informação provinda da *Web* (inicialmente a partir da Wikipedia[10]), que deve melhorar substancialmente os resultados das relações *inclui* e *ocorre_em*; ii) informação morfossintática de aposto e predicativo fornecida pelo analisador PALAVRAS, que de igual forma deve auxiliar principalmente na relação *ident*; iii) utilização de algoritmos de distância de edição, como o utilizado para a correção ortográfica utilizada pelo Google[11], a fim de tratar os casos onde há pequenas diferenças de grafia nos nomes das entidades.

[10] http://pt.wikipedia.org/

[11] http://norvig.com/spell-correct.html

Outro plano para trabalhos futuros é a utilização de ontologias para busca destas informações, e assim melhorar o reconhecimento das relações. Por ora, essa abordagem esbarra na questão da pouca disponibilidade desse tipo de recurso, sendo este de criação e manutenção em geral manual e custosa. Ainda assim, um dos próximos passos deste trabalho é experimentar também esta possibilidade, partindo de um levantamento de ontologias que possam ser utilizadas para o reconhecimento de cada uma das relações apresentadas.

4. Aplicações

Diversas aplicações são possíveis através da utilização do sistema SeRELeP de reconhecimento de relações entre entidades mencionadas. Algumas idéias são relacionadas nesta seção.

Uma aplicação em desenvolvimento é o uso de informação proveniente da relação de identidade no reconhecimento de assuntos populares em conjuntos de notícias num *website*. Este projeto, chamado SeRELeP-Olympics, está em andamento e visa solucionar um problema existente na relação de tópicos mais populares (ou *hot topics*) disposta no portal de notícias Olympicks.net[12], que traz notícias (inclusive com mapeamento geográfico destas), resenhas e *podcasts* sobre as Olimpíadas.

O problema que ocorre no sistema de listagem dos assuntos populares é que estes se baseiam unicamente em informação explícita informada pela fonte da notícia. Notícias provenientes do portal Terra, por exemplo, são marcadas apenas como "Pequim 2008", independentemente de serem notícias referindo-se a protestos, competições ou desportistas específicos. Por esta abordagem, "Pequim 2008" é um assunto popular. Apesar de verídico, é uma informação geral demais, sem especificidade adequada. Idéias de assuntos populares melhor definidos seriam: "César Cielo" (o nadador brasileiro) e "natação", por exemplo. Neste contexto, entendemos que a identificação de entidades e suas relações poderiam aprimorar a identificação desta lista.

Nesta aplicação, além do reconhecimento de EMs, pretende-se reconhecer também outras, que estejam associadas com esporte (etiqueta "sport" do PALAVRAS, como "futebol" ou "natação") e profissões (etiqueta "Hprof", como "técnico" ou "jogador").

Outra aplicação bastante interessante seria a construção automática de ontologias de domínio específico e relacionado ao *corpus* processado pelo SeRELeP, incluindo as EMs constantes nos textos e suas relações.

Além destas, outras aplicações que podem ser associadas a esta tarefa referem-se a sistemas de recomendação de produtos e serviços em *sites* de *e-commerce* (os chamados sistemas de recomendação). Diversos tipos de algoritmos são usados para elencar os itens a serem recomendados para os clientes, e entendemos que o uso de relações semânticas entre entidades pode ser uma abordagem interessante para a questão.

[12] http://olympicks.net

5. Considerações finais

Neste documento, é apresentado um sistema de reconhecimento automático de relações entre EMs em textos da língua portuguesa, fazendo uso de informação lingüística fornecida pelo analisador sintático PALAVRAS.

Este sistema participou do HAREM, uma avaliação conjunta de sistemas com este objetivo, e apresentou resultados bastante promissores. Os resultados são discutidos, assim como são examinadas alternativas para a resolução dos problemas levantados nesta primeira avaliação do sistema.

Como trabalhos futuros, pretende-se utilizar informação provinda da *Web* e outras informações semânticas, tais como ontologias de domínios específicos.

6. Referências e Citações

Bick, Eckhard. (2000). The Parsing System PALAVRAS - Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Tese (Doutorado) – Department of Linguistics, University of Århus, DK..

Linguateca. (2008). Segundo HAREM – Avaliação conjunta de sistemas de Reconhecimento de Entidades Mencionadas. Disponível em: <http://www.linguateca.pt/HAREM/>. Acesso em: setembro/2008.

NIST (*National Institute of Standards and Technology*) e ACE (*Automatic Content Extraction*). (2007). Automatic Content Extraction 2008 Evaluation Plan (ACE08) – Assessment of Detection and Recognition of Entities and Relations Within and Across Documents. Disponível em: <http://www.nist.gov/speech/tests/ace/2008/doc/ace08-evalplan.v1.2d.pdf>. Acesso em: setembro/2008.

Santos, Diana e Cardoso, Nuno (eds.). (2007). Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área. Linguateca. ISBN: 978-989-20-0731-1.

Soon, Wee Meng; Ng, Hwee Tou; Lim, Daniel Chung Yong. (2001). A Machine Learning Approach to Coreference Resolution of Noun Phrases. Computational Linguistics, 27(4):521-544.

Souza, José Guilherme Camargo de. (2007). Resolução automática de correferência aplicada à língua portuguesa. Monografia (Graduação) – Curso de Ciência da Computação, Universidade do Vale do Rio dos Sinos (UNISINOS), BR.

Souza, José Guilherme Camargo de. (2007b) Tiger2XCES: *Software* de conversão de arquivos no formato TigerXML para formato XCES.

Vieira, Renata; Chishman, Rove; Gorziza, Fabiano; Rossoni, Roberta; Rossi, Daniela Rossi; Pinheiro, Clarissa. (2000). Extração de Sintagmas Nominais para o Processamento de Co-referência Nominal. In: Nunes, Maria das Graças Volpe (ed.), V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR). Atibaia, SP. São Paulo: ICMC/USP, pp. 165-173.