Avaliação de Sistemas de Reconhecimento de Entidades Mencionadas

Por: Nuno Francisco Pereira Freire Cardoso
Orientadores: Prof. Mário J. Silva e Prof. Eugénio de Oliveira

Proposta de Tese

Mestrado em Inteligência Artificial e Sistemas Inteligentes
Faculdade de Engenharia da Universidade do Porto

Janeiro de 2006

Conteúdo

1 Introdução

A compreensão computacional da linguagem natural representa um dos maiores desafios da Inteligência Artificial (IA). O desenvolvimento de Sistemas Inteligentes capazes de interagir com os humanos na sua linguagem natural, ou de processar grandes quantidades de informação escrita por humanos, iria beneficiar muitas áreas de investigação dentro da Ciência da Computação.

A avaliação de sistemas de IA é tão antiga quanto a própria área, e é uma etapa fundamental na concretização dos objectivos da IA. Em 1950, Turing propôs um teste para avaliar precisamente o desempenho dos Sistemas Inteligentes em relação ao desempenho feito por um humano, no domínio de compreensão e geração de linguagem natural [Tur50].

O ramo do conhecimento que investiga a interpretação automática de línguas humanas denomina-se processamento de linguagem natural (PLN). A compreensão da linguagem natural é um processo complexo que requer conhecimentos específicos nas suas várias tarefas de interpretação; como tal, o domínio do PLN é decomposto em áreas mais detalhadas que abordam problemas diferentes, como a sumarização de documentos, tradução automática, reconhecimento da fala, resposta a perguntas, extracção de informação (EI) ou o reconhecimento de entidades mencionadas (REM).

A tarefa de REM procura identificar, desambiguar e atribuir um significado semântico a entidades mencionadas (EMs) contidas no texto, como é o caso de nomes de pessoas, de organizações ou de locais. A título ilustrativo, dado o seguinte texto:

```
Eça de Queirós nasceu na Póvoa de Varzim em 1845, e faleceu em
Paris em 1900. Estudou na Universidade de Coimbra.
```

Um Sistema Inteligente em REM poderia detectar e etiquetar as EMs no texto da seguinte forma:

<PESSOA>Eça de Queirós**</PESSOA>** nasceu na **<LOCAL>**Póvoa de Varzim**</LOCAL>** em **<DATA>**1845**</DATA>**, e faleceu em **<LOCAL>**Paris**</LOCAL>** em **<DATA>**1900**</DATA>**. Estudou na **<ORGANIZACAO>**Universidade de Coimbra**</ORGANIZACAO>**.

O REM não é uma tarefa simples, uma vez que as EMs podem ser vagas ou ambíguas, uma característica intrínseca à língua humana [San97, SB04], sendo necessário analisar o contexto onde a EM se insere para definir correctamente a sua categoria semântica.

A decomposição de PLN em problemas que possam ser tratados independentemente permite aferir o progresso obtido em cada um, avaliando as várias aproximações específicas que têm sido propostas para cada caso. No entanto, a especificidade de cada tarefa requer também metodologias de avaliação específicas. No caso particular de REM, a metodologia e o ambiente de avaliação a usar devem ser específicos aos requisitos do problema.

Olhando para o caso português, verifica-se que não havia um plano organizado para acompanhar a evolução dos sistemas de REM sobre textos em português, e medir o seu progresso ao longo do tempo. Surgiu então, no âmbito da Linguateca, a oportunidade de organizar uma avaliação conjunta em SREM que implemente uma nova metodologia de avaliação adequada à problemática da tarefa de REM. A avaliação conjunta tem por objectivo reunir a comunidade científica e concentrar esforços no desenvolvimento de um paradigma de avaliação comum, adoptado e validado por todos os que participam na avaliação, para medir comparativamente os sistemas em torno de tarefas comuns, entre eles, ao longo do tempo.

1.1 Objectivos

O trabalho a apresentar nesta tese deriva da necessidade de planear e organizar avaliações conjuntas específicas para SREMs em português, e de desenvolver uma metodologia de avaliação adequada a REM, abrangendo especificidades não abordadas com profundidade suficiente em eventos de avaliação REM anteriores.

Os objectivos principais do trabalho descrito nesta proposta são:

- Criar uma metodologia para a avaliação de SREMs – a Metodologia HAREM – em conjunto com a comunidade científica interessada em REM. A metodologia será validada durante a organização de um evento de avaliação conjunta em SREM.

- Desenvolver um ambiente de avaliação específico para SREMs – a Plataforma HAREM –, uma bancada de ensaios que meça SREMs que adoptem as regras da Metodologia HAREM, através da geração de valores e gráficos de desempenho. Esta bancada permitirá a qualquer grupo de investigação aplicar as métricas propostas e avaliar comparativamente o seu SREM com outros semelhantes.

- Aplicar e validar a Metodologia HAREM num evento de avaliação conjunta – a Iniciativa HAREM –, usando a Plataforma HAREM. Em conjunto com a comunidade, propõe-se caracterizar o panorama actual de REM em textos escritos em português.

1.2 Resultados

O projecto a desenvolver no âmbito desta tese visa obter os seguintes resultados:

- Uma metodologia nova para avaliação de SREMs, validada pela comunidade científica, que satisfaça as necessidades da comunidade de REM e que servirá de base para medir sistemas de REM e inspirar outras iniciativas semelhantes de avaliação de SREMs.

- Uma colecção de textos ricamente anotada, que constituirá um recurso importante para a melhoria dos sistemas de REM.

- *Software* de avaliação de SREMs, disponível a todos os grupos de investigação que pretendam medir e comparar os seus sistemas. item Uma caracterização do estado da arte em detecção e classificação de EMs por Sistemas

Inteligentes em textos de língua portuguesa, incluindo uma análise crítica das controvérsias geradas no decurso do HAREM.

1.3 Metodologia

Para atingir os objectivos apresentados nesta proposta, propõe-se o desenvolvimento do projecto em três actividades separadas, decorrendo em paralelo e com grande interacção entre si, para concretizar os três objectivos propostos.

A actividade de criação da Metodologia HAREM será feita através dos seguintes passos:

- Elaboração de uma proposta inicial de metodologias de avaliação.
- Disseminação da proposta pelos participantes.
- Melhoramento da proposta, de acordo com as sugestões vindas da comunidade.
- Rectificação da proposta, abrangendo as questões apontadas pela organização e pelos participantes. Nova iteração no processo de disseminação e melhoramento da metodologia.

A actividade de desenvolvimento da Plataforma HAREM seguirá um modelo iterativo:

- Desenho da arquitectura do ambiente de avaliação.
- Especificação das tarefas de cada módulo de *software*.
- Implementação do *software* de acordo com a Metodologia HAREM.
- Validação manual das saídas de cada programa.

A actividade de organização da Iniciativa HAREM segue o seguinte esquema:

- Definição do calendário do evento.

- Implementar a Metodologia HAREM e a Plataforma HAREM segundo os moldes de uma avaliação conjunta.
- Anotação e revisão manual da colecção de textos usada no HAREM - a Colecção Dourada (CD).
- Medição do desempenho dos sistemas de REM, comparando as saídas dos sistemas com a CD.

As actividades propostas serão executadas em paralelo, uma vez que algumas tarefas dependem do progresso feito noutras tarefas.

A interdependência entre as três actividades requer um escalonamento prévio da execução das tarefas, e a complexidade do processo também obriga a um esforço de coordenação da equipa que organiza o HAREM.

A participação activa em eventos de avaliação relacionados também contribuirá para a familiarização com metodologias e ambientes de avaliação existentes.

1.4 Plano de Trabalho

De acordo com a metodologia proposta, o plano de trabalhos proposto divide-se nas seguintes quatro actividades:

Criação da Metodologia HAREM - Reunião da comunidade científica em torno de SREM. Apresentação de uma proposta preliminar e promoção de um debate conjunto, para aperfeiçoar a metodologia, abrangendo todas as questões levantadas durante a construção do ambiente de avaliação e da avaliação conjunta. **Duração:** vinte e três semanas.

Desenvolvimento da Plataforma HAREM - Apresentar uma proposta inicial de arquitectura de avaliação, de acordo com a Metodologia HAREM. Implementação do ambiente de avaliação na Iniciativa HAREM, aperfeiçoamento contínuo do ambiente e do *software* de avaliação, incluindo o desenvolvimento de *software* para a geração de relatórios com estatísticas relativas ao

desempenho global e individual dos sistemas participantes. **Duração:** vinte e três semanas.

Organização da Iniciativa HAREM - Reunião de participantes e observadores interessados em avaliação de SREMs, organização conjunta da Inicativa HAREM e avaliação dos sistemas de REM. **Duração:** vinte e três semanas.

Documentação do HAREM - Exposição da Metodologia, Plataforma e Iniciativa HAREM. Análise dos principais resultados e observações a retirar da avaliação conjunta. Escrita de artigos nas conferências relacionadas e na tese proposta. **Duração:** um ano.

As actividades propostas irão decorrer em paralelo, de acordo com a figura 1.

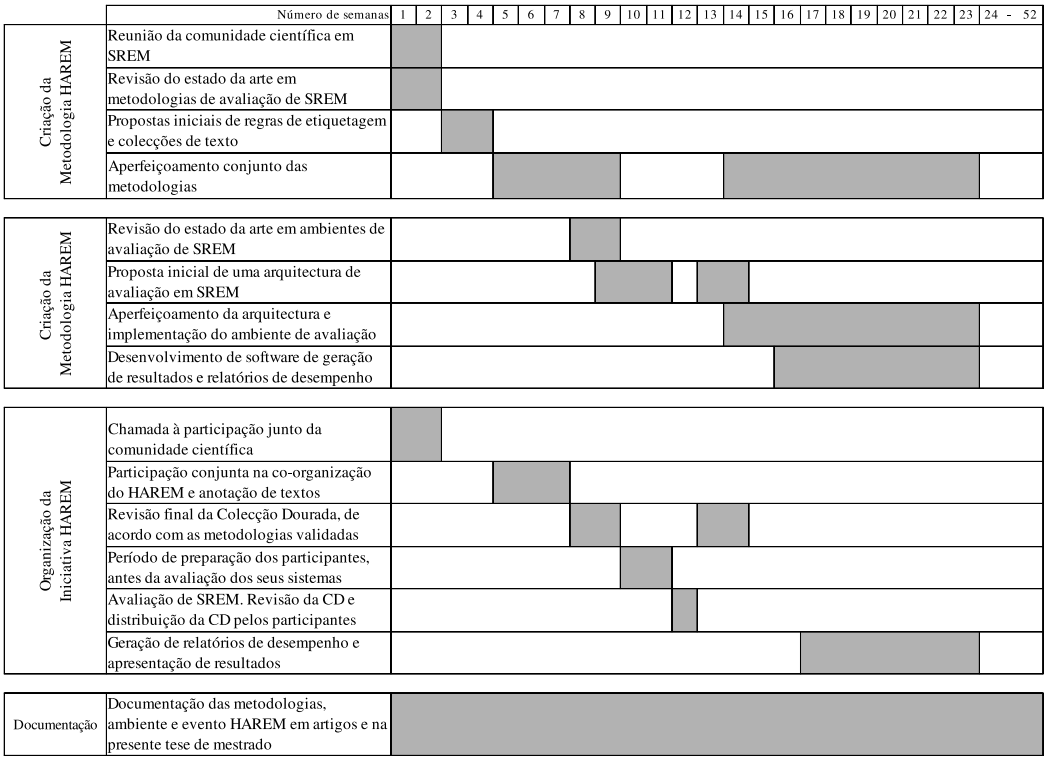

Figura 1: Diagrama de Gantt das quatro actividades da presente proposta de tese.

1.5 Contribuição específica

O trabalho descrito na presente tese foi executado no âmbito da Linguateca, e a organização do HAREM foi fruto de um trabalho de equipa, supervisionado pela Diana Santos. O autor colaborou como co-organizador (com Diana Santos) de todo o processo.

Refira-se a contribuição de Susana Afonso e Anabela Barreiro na construção da colecção dourada, e a contribuição de Nuno Seco e Rui Vilela no desenho e implementação da Plataforma HAREM. Salienta-se também a contribuição de Cristina Mota, na organização do ensaio de 2003 que inspirou o HAREM.

2 Trabalho Relacionado

Um dos objectivos da área de IA é dotar os sistemas computacionais de um comportamento humano nas decisões e acções a tomar no cumprimento de uma determinada função. No entanto, sem métodos de avaliação quantitativa, não é possível seleccionar criteriosamente as soluções mais adequadas de entre as disponíveis para realizar funcionalidades específicas de Sistemas Inteligentes.

A avaliação em PLN desempenha também um papel muito importante, ao medir o progresso feito na compreensão de linguagem natural, por Sistemas Inteligentes. A compreensão da língua humana é uma tarefa complexa e vasta, pelo que a decomposição de PLN em problemas específicos requer avaliações específicas para auxiliar o progresso em cada problema particular. A área de EI é um exemplo de como o aparecimento de inúmeros eventos de avaliação em EI beneficiou bastante o desenvolvimento desta área.

Para fomentar o desenvolvimento na àrea de REM, é necessário desenvolver eventos específicos de avaliação em REM. Esta secção descreve o trabalho feito na criação de metodologias de avaliação, e a sua implementação em conferências de avaliação; de seguida, abordam-se os ambientes de avaliação desenvolvidos até à data, e, finalmente, referem-se alguns eventos importantes de avaliação, na área de REM.

2.1 Metodologias de avaliação

Em 1957, Cleverdon realizou a experiência Cranfield 2 e adoptou as primeiras metodologias próprias de avaliação na área de PLN, no caso particular em EI [Cle67]. Nessa experiência, introduziu-se vários conceitos que ainda hoje são usados na avaliação em EI:

- O uso de uma colecção comum de textos a serem usados por todos os sistemas, para permitir uma avaliação comparativa.
- Um conjunto comum de tópicos, que traduzem necessidades de informação a satisfazer pelos sistemas.
- A noção de relevância de documentos, ou seja, o julgamento humano da importância de um determinado documento para um determinado tópico.
- A introdução das métricas de avaliação precisão e abrangência.

A metodologia iniciada por Cleverdon foi adoptada posteriormente nas primeiras conferências de avaliação de sistemas de EI, que podem ser agrupadas em:

- Avaliações que usam colecções de texto de grandes dimensões nas suas tarefas de avaliação, como o TREC [Har93], o CLEF [PB01] e o NTCIR [KKN$^+$99]. A dimensão das colecções constitui um problema, dado que uma das assumpções da experiência de Cranfield é o conhecimento total da lista de documentos relevantes para cada tópico.

  Como Voorhees constata, é impossível dispor de recursos humanos que consigam avaliar a relevância de todos os documentos, dada a dimensão do corpus usado [Voo01]. A solução passa pela selecção criteriosa de um subconjunto de documentos que, para cada tópico, serão julgados manualmente quanto à sua relevância, enquanto que o resto dos documentos serão considerados irrelevantes para o dado tópico. A esta técnica dá-se o nome de *pooling*.

- Avaliações que usam colecções de texto de pequenas dimensões (os recursos dourados), como o MUC [Hir98],o Parsevel [BAF+91], as Morfolimpíadas [SCR03] ou o HAREM, o que permite a anotação manual das colecções de acordo com as mesmas regras da tarefa para os sistemas. Um dos problemas na criação de recursos dourados é a discordância entre anotadores humanos, um factor importante que é ainda hoje um tema de acesa discussão dentro da comunidade científica de avaliação em EI.

2.2 Ambientes de avaliação

Os resultados obtidos nas avaliações de EI permitem comparar sistemas que implementem estratégias diferentes de resolver uma tarefa comum. A comparação entre sistemas é válida desde que o ambiente de avaliação se mantenha constante, ou seja, a colecção de textos e os tópicos forem os mesmos, e que outros parâmetros de medição se mantenham isolados para evitar a influência de factores que possam alterar as medições; a comparação de resultados entre diversas edições da conferência de avaliação não é válida, uma vez que os tópicos usados são diferentes [Voo01].

Assim sendo, as avaliações precisam de disponibilizar os ambientes de avaliação das suas edições, de maneira a permitir a grupos de investigação a comparação dos seus sistemas num ambiente idêntico ao usado em edições anteriores.

A avaliação na área de PLN aplica métricas derivadas de *Data Mining* e de *Machine Learning*, na medição do desempenho de Sistemas Inteligentes. Gráficos de desempenho, como as *lift charts*, curvas ROC ou curvas de precisão-abrangência, baseiam-se em rácios entre as contagens de decisões correctas (*true positives* e *true negatives*) e decisões incorrectas (*false positives* e *false negatives*) tomadas pelo sistema, uma medida de avaliação usada frequentemente em IA [WF00].

O TREC e o CLEF usam o *software* de avaliação `trec_eval`[1] nos seus ambientes de avaliação, e utilizam curvas de abrangência interpolada vs. precisão para comparar o desempenho dos sistemas. Em 2005, o CLEF melhorou a interface do

[1] `http://www.ir.iit.edu/~dagr/cs529/files/project_files/trec_eval_desc.htm`

ambiente de avaliação, integrando as tarefas de *pooling* e de julgamento humano de relevâncias, com a automatização da recepção de saídas dos participantes e da geração de relatórios de desempenho baseados no `trec_eval` [NF05].

O primeiro ambiente de avaliação na área de REM foi desenvolvido no MUC [Dou98]. Às métricas de precisão, abrangência e medida F, o MUC propôs outras medidas para medir os sistemas, como a Sobre-geração, Sub-geração, Substituição e Erro por Campo de Resposta. A proposta inicial do ambiente de avaliação deste projecto foi inspirado nnas medidas e no *software* de avaliação desenvolvido no MUC.

2.3 Eventos de avaliação em SREM

Conferências MUC

O primeiro MUC teve início em 1987 [SC93], com o intuito de reunir a comunidade de processamento computacional de textos numa avaliação comum. A metodologia usada deriva da experiência de Cranfield, permitindo comparar pela primeira vez vários sistemas em torno de uma tarefa comum de interpretação computacional de textos, usando métricas comuns de avaliação. Citando Gaizauskas, "*if objective measures can be agreed, winning techniques will come to the fore and better technology will emerge more efficiently*" [GHH98].

O tema da primeira edição do MUC foi a interpretação de mensagens no domínio de batalhas navais. Seis sistemas foram valiados, com uma colecção de 10 mensagens. A avaliação foi feita manualmente a partir do resultado de cada sistema, uma vez que o evento não tinha métodos de avaliação definidos.

A 6ª edição do MUC propôs aos participantes, pela primeira vez, uma tarefa de REM [GS96]. A tarefa consistia em identificar e catalogar EMs nas seguintes categorias:

- ENAMEX - compreende as categorias PERSON (pessoa), ORGANIZATION (organização) e LOCATION (local)

- TIMEX - expressões temporais - TIME (hora), DATE (data) e DURATION (período)
- NUMEX - expressões numéricas - MONEY (moeda), MEASURE (medidas), PERCENT (Percentagens) e CARDINAL (quantidades numéricas)

Os resultados obtidos mostram valores de medida-F acima de 0.9 a metade dos sistemas participantes. No entanto, Palmer e Day demonstraram que as EMs de tipo NUMEX e TIMEX são fáceis de detectar, usando um conjunto de regras simples [PD97], e um conjunto significativo de EMs do tipo ENAMEX pode ser detectado, recorrendo a uma lista contendo EMs retirada dos textos de treino e outras EMs mais frequentes. Mikheev et al estudaram a importância da utilização destas listas (ou *gazetteers*) e o impacto destas na tarefa de REM, constatando que é possível criar sistemas eficientes em REM, usando *gazetteers* pequenos [MMG99].

O MUC teve o mérito de ter iniciado a avaliação em áreas como a EI e a REM. Dos eventos, resultaram colecções anotadas (um recurso precioso para diversas comunidades de PLN), métricas simples de avaliação, e uma metodologia específica para avaliação em REM. No entanto, a tarefa de REM não exigiu muito dos sistemas participantes, e ficou a sensação que a avaliação podia ser melhorada.

Conferências pós-MUC

Em 1996 foi organizada a Multilingual Entity Task (MET), a primeira iniciativa multilíngua de avaliação de SREMs. A primeira edição do MET utilizou o inglês e o espanhol como línguas das colecções de texto [MOC96], enquanto que a edição seguinte do MET usou o chinês, o japonês e o inglês nas suas colecções.

O COnLL (Conference on Natural Language Learning) teve o seu início em 1999, para promover a avaliação em áreas específicas de PLN. As edições de 2002 [San02] e de 2003 [SM03] abrangeram o REM e encorajaram a investigação em sistemas REM independentes da língua usada (espanhol e flamengo em 2002, alemão e inglês em 2003).

O ACE [DMP+04] propôs uma tarefa de REM com detecção e classificação das EM e das suas anáforas (nomes, descrições ou pronomes). A colecção incluiu o Inglês, Chinês e o Árabe, e foi disponibilizada em texto, som e imagem. A classificação semântica de EMs abrangeu novas categorias, tais como entidades geo-políticas, armas, veículos e instalações (*facilities*). No entanto, o ACE é um evento que dá mais destaque à tarefa de identificação de coreferências do que à tarefa básica de REM.

O ATIS (Air Travel Information System) [Hir] foi um evento realizado entre 1990 e 1993, para levar a experiência MUC à área de processamento da fala. Ao longo das edições do ATIS foi possível verificar que os sistemas participantes melhoravam o seu desempenho ao longo do tempo, apresentando menores taxa de erro (a métrica principal usada na avaliação), mostrando o papel essencial que os eventos de avaliação possuem no desenvolvimento de uma determinada área de investigação.

O Parseval [BAF+91] foi uma avaliação motivada pela criação do Penn Treebank [MSM93], com o intuito de avaliar os sistemas na interpretação sintática de textos. O Parseval, apesar de não estar directamente relacionada com REM, permite avaliar sistemas na análise de contextos, o que pode contribuir para desambiguar o significado de certas EMs.

Iniciativas da Linguateca

Um dos objectivos da Linguateca é organizar avaliações conjuntas, envolvendo a comunidade científica interessada no processamento computacional da língua portuguesa. A primeira iniciativa de avaliação conjunta organizada pela Linguateca foram as Morfolimpíadas [SCR03], em 2003.

O objectivo das Morfolimpíadas foi a avaliação de analisadores morfológicos; no entanto, o HAREM inspirou-se em alguns procedimentos usados nas Morfolimpíadas.

Após as Morfolimpíadas, a Linguateca alargou o âmbito da sua intervenção para a área de REM e começou a pesquisar novas abordagens à avaliação de

SREM, abrangendo aspectos em REM que não foram considerados nas avaliações em REM anteriores (como a vagueza e a ambiguidade das EMs) e fornecer dados para a construção de uma nova metodologia de avaliação em REM (o que aconteceu no HAREM, o evento apresentado neste trabalho).

A iniciativa colocou em causa a hierarquia rígida de categorias e de regras de etiquetagem propostas nas anteriores avaliações, argumentando que a nova metodologia em avaliação de REM tem de ser desenvolvida em conjunto com a comunidade científica na área de REM. Para o efeito, foi realizado um estudo com nove investigadores (interessados em participar num futuro evento de avaliação em REM), que anotaram livremente 20 extractos das colecções CETEM-Público [SR01] e CETENFolha.

Os investigadores detectaram entre 179 a 250 EMs, e o nível de concordância entre anotações foi de 46% (no caso de EMs de nomes próprios). Este valor de concordância, bem como a diversidade de categorias e sub-categorias semânticas usadas pelos participantes para classificar as EMs, demonstram que as avaliações anteriores em REM não reproduzem correctamente a problemática de detecção e classificação de EMs, sendo necessário discutir em conjunto e procurar um consenso quanto à metodologia a usar e às directivas a adoptar, antes de organizar um evento de avaliação conjunta em REM.

3 Conclusão

O projecto proposto já se encontra na sua fase de documentação. As actividades de desenvolvimento de metodologias, implementação de ambiente de avaliação e o evento de avaliação conjunta já foram concluídas.

4 Glossário

Avaliação Conjunta - Evento de avaliação realizado por avaliadores independentes, com a colaboração activa dos participantes na definição dos moldes da avaliação e na criação do ambiente de avaliação.

CLEF - Cross Language Evaluation Forum - Evento de avaliação de sistemas de EI em línguas europeias, em contextos monolngua e multilngua.

CD - Colecção Dourada - Colecção de textos contendo anotações manuais das EMs nela contidas, e revista por diversas pessoas, de acordo com um conjunto de regras de etiquetagem. A avaliação em SREM é feita ao comparar a colecção dourada com as saídas dos sistemas de REM.

EM - Entidade Mencionada - Tradução de *Named Entity (NE)*, é o conjunto de termos usados no texto para representar uma determinada entidade, com um forte significado semântico.

HAREM - projecto realizado pela Linguateca em 2004-2005, com o objectivo de criar uma Avaliação Conjunta de Sistemas de Reconhecimento de Entidades Mencionadas em textos escritos em português.

NTCIR - NII-NACSIS Test Collection for IR Systems - Evento de avaliação de sistemas e tecnologias na área de EI, para línguas asiáticas.

Penn Treebank - Colecção de textos em inglês com anotação sintática estruturada.

REM - Reconhecimento de Entidades Mencionadas - Tradução de *Named Entity Recognition (NER)*, representa a área específica de Extracção de Informação que detecta referências a Entidades Mencionadas no texto, e que regista o seu sentido semântico.

SREM - Sistema de Reconhecimento de Entidades Mencionadas - Sistema computacional de REM que implementa técnicas de PLN para anotação de EMs no texto, recorrendo a etiquetas.

TREC - Text REtrieval Conference - Conferência co-patrocinada pelo NIST e pelo ARDA, com o objectivo de promover a investigação na área de EI em grandes colecções de texto, e avaliar as metodologias implementadas.

Referências

[BAF+91] E. Black, S. Abney, D. Flickenger, C. Gdaniec, R. Grishman, P. Harrison, D. Hindle, R. Ingria, F: Jelinek, J: Klavans, M. Liberman, M. Marcus, S. Roukos, B. Santorini e T. Strzalkowski. A procedure for quantitatively comparing the syntactic coverage of English grammars. Em *Proceedings of the February 1991 DARPA Speech and Natural Language Workshop*, páginas 306–311, Pacific Grove, CA, Fevereiro 1991.

[Cle67] Cyril W. Cleverdon. The Cranfield tests on index language devices. *Aslib Proceedings*, 19(6):173–193, 1967.

[DMP+04] George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel e Ralph Weischedel. The Automatic Content Extraction (ACE) Program. Tasks, Data and Evaluation. Em Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa e Raquel Silva, editores, *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC*, páginas 837–840, Lisboa, Portugal, 26-28 Maio 2004.

[Dou98] A. Douthat. The Message Understanding Conference Scoring Software User's Manual. Em *Proceedings of the 7th Message Understanding Conference MUC-7*, 1998. `http://www-nlpir.nist.gov/related_projects/muc/muc_sw/muc_sw_manual.htm%l`.

[GHH98] R. Gaizauskas, M. Hepple e C. Huyck. A scheme for comparative Evaluation of Diverse Parsing Systems. Em *Proceedings of the 1st International Conference on Language Resources and Evaluation*, Granada, Espanha, 1998.

[GS96] Ralph Grisham e Beth Sundheim. Message Understaning Conference - 6: A Brief History. Em *Proceedings of the 16th International Con-*

ference on Computational Linguistics (COLING-96), páginas 466–471, Copenhagen, Danmark, 1996.

[Har93] Donna Harman. Overview of the first TREC conference. Em *SIGIR '93: Proceedings of the 16th annual international ACM SIGIR conference on Research and Development in Information Retrieval*, páginas 36–47, Nova Iorque, EUA, 1993. ACM Press.

[Hir] Lynette Hirschman. Evaluating spoken language interaction: Experiences from the DARPA spoken language program 1980-1985. Em S. Luperfoy, editor, *Spoken Language Discourse*. MIT Press, CAmbridge, MA.

[Hir98] Lynette Hirschman. The evolution of Evaluation: Lessons from the Message Understanding Conferences. *Computer Speech and Language*, 12(4):281–305, 1998.

[KKN+99] N. Kando, K. Kuriyama, T. Nozue, K. Eguchi, H. Kato, e S. Hidaka. Overview of IR Tasks at the First NTCIR Workshop. Em *Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, páginas 11–44, Tóquio, Japão, Agosto 1999.

[MMG99] Andrei Mikheev, Marc Moens e Claire Grover. Named Entity recognition without Gazetteers. Em *Proceedings of EACL'99*, páginas 1–8, Bergen, Norway, 8-12 Junho 1999.

[MOC96] Roberta Merchant, Mary Ellen Okurowski e Nancy Chinchor. The Multilingual Entity Task (met) overview. Em *Proceedings of TIPSTER Text Program (Phase II)*, 1996.

[MSM93] M. Marcus, B. Santorini e M. Marcinkiewicz. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, (19):313–330, 1993.

[NF05] Giorgio Di Nunzio e Nicola Ferro. DIRECT: a System for Evaluating Information Access Components of Digital Libraries. Em *Proceedings of the 9th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2005)*, 18-23 Setembro 2005.

[PB01] Carol Peters e Martin Braschler. European Research Letter: cross-language system evaluation: the CLEF campaigns. *Journal of the American Society for Information Science and Technology*, 52(12):1067–1072, 2001.

[PD97] D. Palmer e D. Day. A statistical profile of the named entity task. Em *Proceedings of Fifth ACL Conference for Applied Natural Language Processing (ANLP-97)*, Washington D.C., EUA, 1997.

[San97] Diana Santos. The importance of vagueness in translation: Examples from english to portuguese. *Romansk Forum*, 5:43–69, Junho 1997.

[San02] Erik F. Tjong Kim Sang. Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. Em *Proceedings of CoNLL-2002*, páginas 155–158, Taipei, Taiwan, 2002.

[SB04] Diana Santos e Anabela Barreiro. On the problems of creating a consensual golden standard of inflected forms in portuguese. Em In Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa e Raquel Silva, editores, *Proceedings of LREC'2004, Fourth International Conference on Language Resources and Evaluation*, 26-28 Maio 2004.

[SC93] Beth Sundheim e N. Chinchor. Survey of the Message Understanding Conferences. Em *Proceedings of the Human Language Technology Conference*, páginas 56–65, Princeton, NJ, Março 1993.

[SCR03] Diana Santos, Luís Costa e Paulo Rocha. Cooperatively evaluating Portuguese morphology. Em Nuno J. Mamede, Jorge Baptista, Isabel Trancoso e Maria das Graças Volpe Nunes, editores, *Computational Processing of the Portuguese Language, 6th International Workshop, PROPOR 2003*, páginas 259–266, Faro, Portugal, Junho 2003. Springer Verlag.

[SM03] Erik F. Tjong Kim Sang e Fien De Meulder. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. Em *Proceedings of CoNLL-2003*, páginas 142–147, Edmonton, Canadá, 2003.

[SR01] Diana Santos e Paulo Rocha. Evaluating cetempúblico, a Free Resource for Portuguese. Em *Meeting of the Association for Computational Linguistics - ACL*, páginas 442–449, Toulouse, 9-11 Julho 2001.

[Tur50] Alan M. Turing. Computing Machinery and Intelligence. *Mind*, 59:433–460, 1950.

[Voo01] Ellen Voorhees. The Philosophy of Information Retrieval Evaluation. Em *Proceedings of the 2nd Workshop of the Cross-Language Evaluation Forum, CLEF 2001*, Darmstadt, Alemanha, 3-4 Setembro 2001.

[WF00] Ian H. Witten e Eibe Frank. *Data Mining - Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, 2000.