REMando para o futuro: reconhecimento de entidades mencionadas e não só

Diana Santos e Nuno Cardoso

Linguateca
www.linguateca.pt

Resumo

- Introdução
  - O que é REM?
  - Para que serve?
  - Breve história
- REM em português
  - Introdução ao HAREM
  - Colecções do HAREM
  - Resultados no HAREM

Resumo (continuação)

- Dentro de um sistema de REM
  - Sistemas de REM em português
  - Dissecando o REMBRANDT
- No meio do futuro: o ReRelEM
- Futuro do REM
  - REMBRANDT: Explorar os recursos emergentes (Wikipédia, Dbpedia)
  - GikiCLEF
  - A relação entre o REM e o PLN em geral

O que é REM?

- **R**econhecimento de **E**ntidades **M**encionadas: identificação e classificação de entidades (designadas/representadas por nomes próprios) no texto.

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

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

Categorias semânticas:
Cidade, **Ano**, **Pessoa**, **Universidade**

Para que serve REM?

Tarefa relevante em diversas áreas de PLN:

- **Tradução automática** ("Rio de Janeiro" ≠ "January River")
- **Resposta automática a perguntas** ("Quem é Castelo Branco?" ≠ "Onde fica Castelo Branco?")
- **Análise sintáctica** (castelo/{nome} branco/{adj} vs castelo branco/{nome_proprio})
- **Desambiguação de pesquisas** com âmbito geográfico (romances de Castelo Branco vs turismo em Castelo Branco)
- **Sumarização de textos** (por exemplo para resolução de anáforas)
- etc.

REM é fácil?

- Para algumas entidades mencionadas (EM), sim.
  - "O arquitecto Silva desenhou o projecto."
- Para outras, é bem complicado...
  - Identificação:
    - Portugal e Espanha assinaram um acordo...
    - Em São Tomé e Príncipe assinaram um acordo...
  - Vagueza / ambiguidade na identificação:
    - Governo de Cavaco Silva – uma ou duas EM?
    - **u**niversidade de Lisboa vs **U**niversidade de Lisboa...
  - Classificação:
    - Ajudem os Bombeiros! - Organização ou grupo de pessoas?
    - Sentido mais comum da EM pode não ser o que se pretende no contexto

Vários sentidos da EM "Portugal"

- Portugal não é só um país...
  - **Portugal** venceu a Alemanha por 3-0.
  - **Portugal** votou 'não' na ONU.
  - O meu querido **Portugal** da infância...
  - **Portugal** tem muitas praias.
  - (João) **Portugal** canta hoje em Lisboa.
  - As acções da **Portugal** Telecom S.A...

Breve história do REM

Breve história do REM

- 1995: Primeira iniciativa de avaliação separada na 6ª edição do Message Understanding Conference (MUC-6)
  - **Tarefa**: Reconhecer ENUMEX (PERSON, LOCATION, ORGANIZATION), NUMEX (MONEY, PERCENT) e TIMEX (TIME, DATE)
  - Colecção de **30** textos (MUC-6) e **100** textos (MUC-7), retirados de uma colecção de 58.000 documentos do Wall Street Journal (Jan/93 a Jun/94)
  - Primeiras métricas de avaliação de sistemas de REM
  - Análise para determinar a relevância estatística das diferenças de desempenho entre os sistemas participantes

Breve história do REM (cont.)

- 2000- : ACE
  - Tarefa: *EDT – Entity Detection and Tracking*
  - Inclui resolução de anáforas
  - Categorias abrangem o domínio militar (armas, veículos, instalações, etc)
  - Textos em inglês, chinês e árabe
  - Colecções de texto, som e imagem
- 2002 e 2003: CoNLL
  - Classificação: PER, LOC, ORG, MISC (pessoas, locais, organizações, miscelânea)
  - Textos (~200) em espanhol e flamengo (2002), alemão e inglês (2003)
  - Identificação já feita previamente na colecção

Breve história do REM (continuação)

Pontos positivos do REM do MUC:

- Despoletou avaliações noutras línguas (ex: MET - espanhol, chinês, japonês)
- REM estabelecido como uma tarefa essencial genérica de EI (extracção de inf)
- Resultado da maioria dos sistemas: medida F > 90%

Pontos negativos:

- Categorias mais frequentes, mas não motivadas pela língua
- Dificuldades reais de REM (ex: vagueza) "passsaram ao lado"
- Colecção usada: só textos jornalísticos...
- Tarefa de REM "resolvida" segundo os resultados dos participantes, ou a tarefa era "fácil" demais?

Antes do HAREM, reflexão: Como seria uma avaliação de REM para o português? Diferente? Mais do mesmo? Os resultados seriam os mesmos?

HAREM

- HAREM - Avaliação conjunta de sistemas de REM, organizada pela Linguateca
- Modelo do HAREM: Avaliação conjunta
  - participantes e interessados ajudaram na definição da tarefa de avaliação
  - anotaram as colecções de treino, sugeriram categorias e tipos
  - a grande maioria participou no HAREM

O imapcto do HAREM

- Antes do HAREM: que sistemas existentes para o português?
- Depois do HAREM: temos conhecimento de 17 sistemas
  (só dois não participaram no HAREM)
  - CaGE,
  - Cortex,
  - F-EXT-WS
  - LX-NER,
  - Malinche,
  - NERUA,
  - PALAVRAS_NER,
  - PorTexTO,
  - Priberam,
  - REMBRANDT,
  - REMMA,
  - RENA,
  - SIEMÊS,
  - SEI-Geo,
  - SeRELeP,
  - SMELL,
  - Stencil/NooJ,
  - XIP

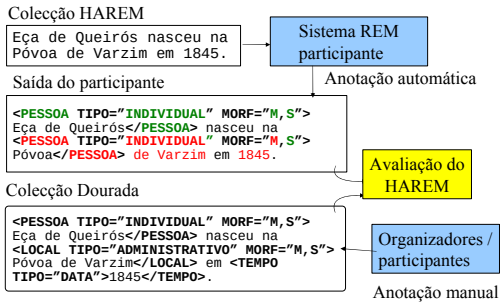
Avaliações no HAREM

- Primeiro HAREM: 2 eventos:
  - Primeiro evento: Fevereiro de 2005 (10 participantes, 6 países)
  - Segundo evento, "Mini-HAREM": Abril de 2006 (5 participantes)
  - 3 tarefas:
    - Identificação
    - Classificação Semântica
    - Classificação morfológica
- Segundo HAREM: 1 evento
  - Abril de 2008 (10 Participantes, 2/4 países)
  - 3 tarefas:
    - HAREM "clássico": Identificação + Classificação,
    - TEMPO: Normalização temporal
    - ReRelEM: Detecção de relações entre EM

Características do HAREM

- Leque de categorias obtidas "de baixo para cima" por vários anotadores (vs categorias pré-determinadas)
- 10 categorias (PESSOA, LOCAL, ORGANIZACAO, TEMPO, ABSTRACCAO, COISA, OBRA, ACONTECIMENTO, VARIADO) com 41 tipos (bem mais detalhada que o MUC/CoNLL)
- Uso de uma **Colecção HAREM** (CH) com ~1200 documentos, de várias origens, géneros variantes de português (vs textos exclusivamente jornalísticos)
- **Colecção dourada** (CD), ~1/8 da CH, exaustivamente anotada e revista por vários anotadores
- Vagueza na identificação e classificação de EM tomada em conta: nada de decisões arbitrárias

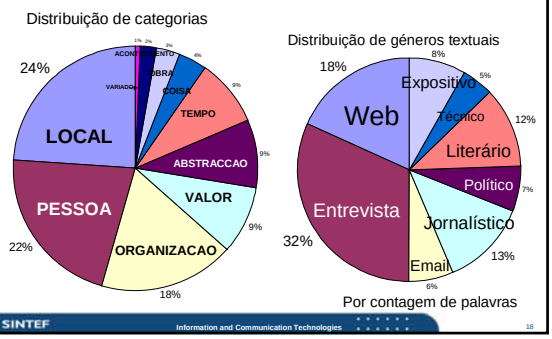
Metodologia HAREM



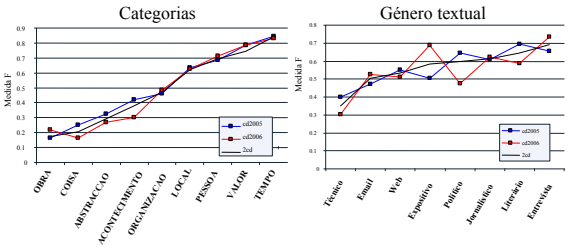
Colecção Dourada (CD)

- Excertos da Colecção HAREM (~1/8) com as respectivas EM anotadas manualmente pelos participantes e pela organização
- Vários géneros textuais (jornalístico, oral, web, correio electrónico, literário, expositivo, técnico, político)
- Várias variantes (português de Portugal, português brasileiro, textos de origem timorense, angolana, macaense e moçambicana, Wikipédia)
- Primeiro HAREM teve duas CD (~130 docs cada)
- Segundo HAREM: Colecção HAREM nova, CD para a pista clássica (129 docs), TEMPO (30 docs) e ReRelEM (12 docs)

Colecção Dourada (cont.)

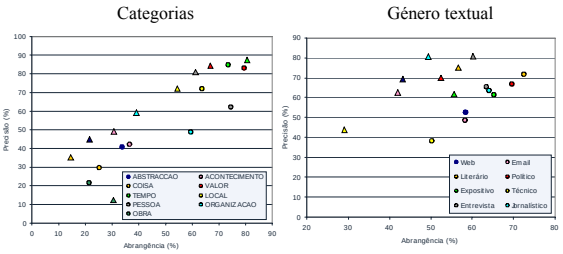


Desempenho no Primeiro HAREM



Resultados para o MiniHAREM, tarefa de classificação semântica

Desempenho no Primeiro HAREM



Resultados para o MiniHAREM, tarefa de classificação semântica, para as 2 CD.

Dentro dos sistemas de REM

Sistemas REM 1: baseados em regras (gramáticas)

- PALAVRAS_NER:
  - Regras gramaticais, gramática dependencial, do PALAVRAS (~6000 regras)
  - Almanaques de 70.000 lexemas + 17.000 EM pré-classificadas.
  - Hierarquia interna de 6 categorias e 18 tipos
  - Melhor resultado no primeiro evento do Primeiro HAREM
- XIP
  - Xerox Incremental Parser, um analisador sintáctico dependencial
  - Uso de regras gramaticais específicas para o português no XIP
  - Atento ao contexto onde a EM se insere

Sistemas REM 2: Aprendizagem automática

- Cortex
  - Aprendizagem automática para reconhecimento de padrões
  - Inclui módulos dependentes da língua e regras manuais
  - Sem almanaques, usa um grupo inicial de lexemas, vai aprendendo novos lexemas com texto novo.
  - Eficiência depende no nº de textos novos alimentados ao Cortex
  - Melhor resultado no Mini-HAREM
- NERUA
  - Adaptação de um sistema espanhol de REM com algoritmos de aprendizagem automática
  - Treinado com um pedaço da Colecção HAREM anotada e distribuído aos participantes

Sistemas REM 3: Almanaques

- SIEMÊS
  - Usa 5 regras de associação dos termos das EM no seu almanaque
  - Um almanaque extenso de EM (REPENTINO), com 450.000 entradas – o maior almanaque reportado por um sistema REM em português.
  - Hierarquia de 11 categorias e 102 tipos
- Stencil/NooJ
  - Dicionários de palavras-chave e gramáticas locais.
  - **Não utiliza almanaques** – só regras.

Sistemas REM 4: Uso de ontologias

- CaGE
  - Usa uma ontologia geográfica (Geo-Net-PT 01) + regras gramáticas básicas para o reconhecimento único de nomes geográficos.
- Sei-Geo
  - Sistema dedicado à extracção de nomes geográficos e detecção de relações entre EM LOCAL, com o objectivo enriquecer ontologias geográficas com novas entidades devidamente validadas.
- Priberam
  - Ontologia multilingue proprietária, com base em relações semânticas e conceptuais entre palavras e expressões
  - Sistema de regras gramáticas manuais, usando anotações morfossintácticas.
  - Melhor resultado do Segundo HAREM

Sistemas REM 5: Explorando a Wikipédia

- REMMA:
  - Focado em textos médicos
  - Análise da primeira frase/parágrafo da Wikipédia para a classificação das EM
- REMBRANDT:
  - Categorias da Wikipédia usadas na classificação de EM
  - Sem análise morfossintáctica

Dissecando um sistema de REM: o REMBRANDT

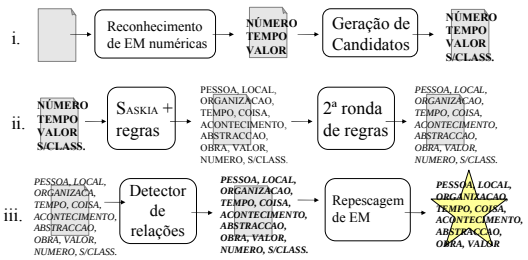
Decisões a tomar ao construir um sistema REM

- Dedicado a uma língua, ou independente da língua?
- Regras gramáticas inseridas manualmente, ou aprender regras de forma automática?
- Que almanaques usar? Qual o tamanho dos almanaques? Construo os meus almanaques, ou uso algum recurso público?
- Usar textos de treino para ensinar o sistema?
- Fazer a identificação e depois a classificação, ou as duas em simultâneo?
- Preferências por um sistema rápido, ou completo (caso não se consiga os dois)? Abrangente ou preciso?

REMBRANDT

- Objectivo:
  - reconhecer nomes geográficos no texto, para posteriormente detectar o âmbito geográfico do documento
  - Processar grandes quantidades de texto de forma rápida
- Mas, para tal, é preciso distinguir o verdadeiro sentido da EM (ver o exemplo de Portugal), e saber quando não são EM geográficas
- Estratégia usada:
  - reconhecer **TODAS** as EM, para facilitar a desambiguação
  - Wikipédia como fonte de informação:
    - Páginas estruturadas e categorizadas, abrange todos os tipos de EM
    - Páginas de desambiguação, identificadores únicos para cada conceito
  - Usar regras gramáticas específicas para o português, repescar EM com relações capturadas

Receita do REMBRANDT



Explorando a Wikipédia (1)

- Categorias Wikipédia: *Cantores de Portugal*, *Cantores de heavy metal*
- Cantores ➔ PESSOA INDIVIDUAL

Explorando a Wikipédia (2)

- Páginas de desambiguação:
  - Ligações às páginas das entidades com diferentes significados, mas mesma EM
  - Resolução de acrónimos

Explorando a Wikipédia (3)

- Desambiguação de EM:
  - Ex: "Armstrong participou activamente no projecto Gemini"
  - Qual é o Armstrong certo? Local? Pessoa? Qual deles?
  - Página de **Neil** Armstrong tem uma referência ao projecto Gemini – boa pista de desambiguação

Prós / Contras da estratégia REMBRANDT

Prós:
- Actualização da Wikipédia → actualizações do almanaque
- Ainda com poucas regras (~200), comparando p.ex. ao PALAVRAS. Mais regras deve melhorar precisão / abrangência
- Páginas de desambiguação da Wikipédia facilitam a desambiguação de EM e a sua referenciação em entidades diferentes

Contras:
- Manutenção / encadeamento das regras complica-se cada vez mais
- Gestão de conflitos de regras cada vez mais complicado
- "Identificação, depois classificação" não é a melhor opção em muitas EM complexas – "gerador de candidatos a EM" muito ingénuo
- Detector de relações lento para documentos grandes
- Categorias Wikipédia por vezes caóticas...

Estado do REMBRANDT

- http://xldb.di.fc.ul.pt/Rembrandt/:
  - Serviço de rede (limitado)
  - Pacotes para descarregar, documentação, ajuda na instalação e configuração
  - Promover o retorno de problemas / sugestões / comentários
  - Adaptação ao inglês, uso da DBpedia (mais preciso e rápido)
- Futuro:
  - Adicionar resolução de topónimos
  - Mais ênfase na DRE, para melhor capacidade de desambiguação
  - Anotar o corpo dos textos da Wikipédia toda, para que também possam ser usados em EI (lição nº 1 do GikiCLEF)
  - Tentar ser o "Corpógrafo" das EM, uma bancada completa de anotação automática e rectificação manual de documentos

No meio do futuro? O ReRelEM

Cristina Mota, Cláudia Freitas, Paula Carvalho, Hugo Oliveira, Diana Santos

ReRelEM

- Tarefa de detecção de relações entre EM, do Segundo HAREM
- Objectivo: detectar relações:
  - Identidade
  - Inclusão (inclui, incluido)
  - Ocorre em, Sede de
  - Outra

ReRelEM – o que anotar?

Identidade

… poderia ser resolvida com um simples teste de **ADN** (**DNA**).

Inclusão (inclui / incluido)

Lobos recebidos em apoteose
(…), o capitão **Vasco Uva** explicou por que houve uma empatia tão grande entre …

Localização (ocorre_em / sede_de)

GP Brasil
Não faltou emoção (…) no **Circuito José Carlos Pace** desde a primeira volta…

Outra (outras relações relevantes)

ReRelEM – como anotar?

ID da EM relacionada

tipo de relação entre EM

Atributos HAREM + **COREL** e **TIPOREL**

Um dos telescópios já está pronto e em funcionamento no <EM ID="**L111**">Havaí</EM>, <EM ID="L165" COREL="**L111**" TIPOREL="inclui">EUA</EM>

Relações entre EM vagas – relações entre facetas

contra <EM ID="h-11" CATEG="**PESSOA**" TIPO="**INDIVIDUAL**">**Diogo Barbosa Machado**</EM> na <EM ID="h-64" CATEG="**LOCAL|OBRA**" TIPO="**VIRTUAL|REPRODUZIDA**" COREL="h-11" TIPOREL="QUEM^^outra^^hum-49949-11^^PESSOA">**Bibliotheca Lusitana**</EM>

HAREM → ReRelEM

Portugal perdeu para a **Alemanha** nas quartas de final da **Eurocopa**. Vi o jogo na **Praça da República**, e mesmo com a derrota os bares de **Coimbra** estavam cheios.

HAREM → ReRelEM

Portugal

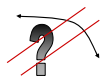
Coimbra

<EM ID="h-37" CATEG="**PESSOA**" TIPO="**GRUPOMEMBRO**">**Portugal**</EM> perdeu para a <EM ID="h-38" CATEG="**PESSOA**" TIPO="**GRUPOMEMBRO**">**Alemanha**</EM> na <EM ID="h-39" CATEG="**ACONTECIMENTO**" TIPO="**ORGANIZADO**">**Eurocopa**</EM>. Vi o jogo na <EM ID="h-40" CATEG="**LOCAL**" TIPO="**HUMANO**" SUBTIPO="**RUA**">**Praça da República**</EM>, e mesmo com a derrota os bares de <EM ID="h-41" CATEG="**LOCAL**" TIPO="**HUMANO**" SUBTIPO="**DIVISAO**">**Coimbra**</EM> estavam cheios.

HAREM → ReRelEM

Portugal

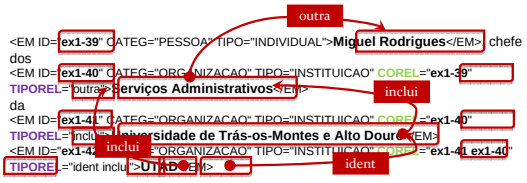
Coimbra



<EM ID="h-37" CATEG="**PESSOA**" TIPO="**GRUPOMEMBRO**">**Portugal**</EM>

<EM ID="h-41" CATEG="**LOCAL**" TIPO="**HUMANO**" SUBTIPO="**DIVISAO**">**Coimbra**</EM>

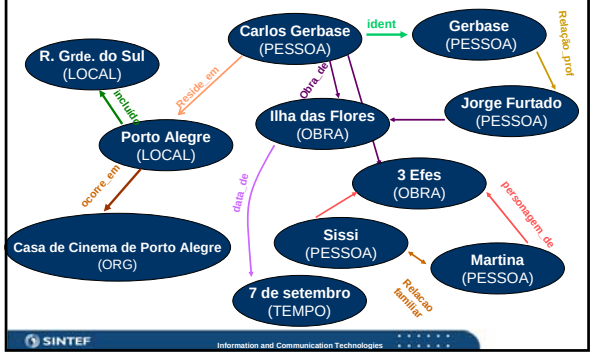
Esquema de anotação



- Não é preciso anotar todas as relações
- Os programas de avaliação fazem-no

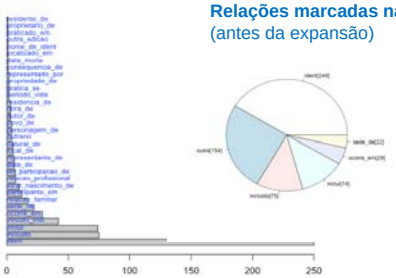
Vilain et al. (1995)

A ident B ∧ B ident C	⇒ A ident C
A inclui B ∧ B inclui C	⇒ A inclui C
A inclui B ∧ B sede_de C	⇒ A sede_de C
A ident B ∧ any_rel C	⇒ A any_rel C



CD ReRelEM (3)

Relações marcadas na CD
(antes da expansão)



Futuro do REM, ou: para onde vamos?

O REM está resolvido? Chegámos a algum lado? Vale a pena continuar a investir nesta área, ou avançar para outras?

Do ponto de vista do REMBRANDT

- Explorar as relações entre EM é importante – o ReRelEM mostrou isso
- Apostar mais em almanaques "dinâmicos" - Wikipédia – do que em almanaques internos
- Referenciar as entidades desambiguadas com identificadores únicos (ex: DBpedia)

A nível internacional, ainda se fala de REM?

Ainda, mas a um nível diferente

- Multilingue (detectar nomes próprios em texto árabe, chinês, etc.)
- Na rede (ou seja, aplicado a motores de busca como o Google)
  - Em que o problema são demasiadas facetas de cada EM
- Integrado em aplicações maiores
  - Sistemas de detecção de tópicos ou portais
  - Sistemas de RAP ou RIG
- Tratando da metonímia sistemática
- Tratando de tipos de texto especiais
  - Genética, medicina, engenharia

Lack of agreement on the purpose of a discipline: what is QA?

Wilks (2005:277):

providing ranked answers [...] is quite counterintuitive to anyone taking a common view of questions and answers. "Who composed Eugene Onegin? and the expected answer was Tchaikowsky [...] listing Gorbatchev, Glazunov etc. is no help

Karen Sparck-Jones (2003):

Who wrote "The antiquary?"

The author of Waverley

Walter Scott

Sir Walter Scott

Who is John Sulston?

Former director of the Sanger Institute

Nobel laureate for medicine 2002

Nematode genome man

There are no context-independent grounds for choosing any one of these

ESSLLI 2007

Duas visões de RAP

A visão da RI: extracção de passagens antes da IE

Mas: "what colour is the sky?" passages with *colour* and *sky* may not have *blue* (Roberts & Gaizauskas, 2003)

A visão da IA: compreensão profunda

Mas: "where is the Taj Mahal?" (Voorhees & Tice, 2000) how can you know what the user has in mind/knows, and therefore wants to know?

De facto, as coisas fáceis para uma das abordagens são as difíceis para a outra

ESSLLI 2007

GikiCLEF

Perguntas "abertas" à Wikipédia (não se sabe a priori quantas respostas certas)

Uma tarefa difícil tanto para pessoas como para computadores

Uma situação realística em que recolha cruzada e multilingue são necessárias ou pelo menos úteis

A decorrer na edição deste ano do CLEF, CLEF 2009, **http://www.linguateca.pt/GikiCLEF**

Mistura de RAP e de RIG, continuando o GikiP, uma AC piloto

10 colecções, 50 tópicos motivados culturalmente

Topic titles in GikiP 2008: cultures

ID	English topic title
GP1	Which waterfalls are used in the film "The Last of the Mohicans"?
GP2	Which Vienna circle members or visitors were born outside the Austria-Hungarian empire or Germany?
GP3	Portuguese rivers that flow through cities with more than 150,000 inhabitants
GP4	Which Swiss cantons border Germany?
GP5	Name all wars that occurred on Greek soil.
GP6	Which Australian mountains are higher than 2000 m?
GP7	African capitals with a population of two million inhabitants or more
GP8	Suspension bridges in Brasil
GP9	Composers of Renaissance music born in Germany
GP10	Polynesian islands with more than 5,000 inhabitants
GP11	Which plays of Shakespeare take place in an Italian setting?
GP12	Places where Goethe lived
GP13	Which navigable rivers in Afghanistan are longer than 1000 km?
GP14	Brazilian architects who designed buildings in Europe
GP15	French bridges which were in construction between 1980 and 1990

Exemplos de sistemas GikiCLEF em uso...

Perdidos nas montanhas norueguesa, uma família italiana...

Em que países da Europa se usa bidé?

- Portugal
- Espanha
- Andorra
- Itália
- ...

Exemplos de sistemas GikiCLEF em uso...

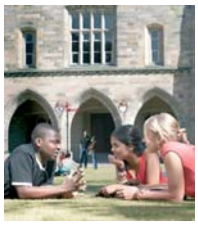
No meio de um desafio de futebol a passar numa estação estrangeira de televisão...

Que países sul americanos usam amarelo nos seus uniformes?

- Brasil
- Nicarágua

Exemplos de sistemas GikiCLEF em uso...

Um encontro de jovens alemães (ou um bate-papo), discutindo: aonde havemos de ir estudar mantendo o contacto?



Que cidades alemães têm mais de uma universidade?

- Berlim
- Bona
- Colónia
- Aachen
- Bremen
- Ausgsburgo
- Hamburgo
- ...

Exemplos de sistemas GikiCLEF em uso...

Em que museu americano existia aquela obra do Picasso sobre a qual vimos aquele programa muito interessante no Natal?

Que museus americanos têm obras de Picasso?

- Museum of Modern Art
- Museum of Fine Arts (Boston)
- Solomon R. Guggenheim Museum
- Metropolitan Museum of Art
- National Gallery of Art
- Art Institute of Chicago
- Denver Art Museum
- (...)

Exemplos de sistemas GikiCLEF em uso...

Que poetas romenos influenciaram Mircea Eliade? Investigação preliminar para um mestrado em literatura romena, antes de ir levantar os livros à biblioteca

Que poetas romenos publicaram livros de baladas antes de 1931?


Dimitrie Bolintineanu


Vasile Alecsandri


George Coşbuc


Elisabeta de Neuwied

REMando para o futuro...

Anotação dos corpos linguísticos com EM já é uma realidade

O REM não se pode desligar de tudo o resto

O REM é apenas a "parte mais fácil" da compreensão alargada e robusta de texto

Muito se aprende começando pelo mais fácil, desde que não se faça hipóteses redutoras ou simplesmente erradas, e não se deixe de pensar nos problemas / tarefas que se quer resolver

Não há nada de mágico nas maiúsculas... Ou há?

Todos os problemas da língua se podem estudar com base em EM... todas as áreas da linguística têm/precisam de ter as EM em consideração