Avaliação Conjunta de Sistemas de Reconhecimento de Entidades Mencionadas

Cristina Mota
cristina@label.ist.utl.pt

Linguateca, Pólo do LabEL, IST
http://label.ist.utl.pt
http://www.linguateca.pt

AvalON'2003
Encontro de Avaliação Conjunta de Sistemas de Processamento Computacional do Português
Faro, Portugal – 26 de Junho de 2003

Qual é o objecto de avaliação?

Sistemas ou módulos de sistemas que façam
Reconhecimento de Entidades Mencionadas (Named Entity Recognition)

Nomes próprios

Pessoas
- Fernando Pessoa
- Maria do Carmo Sampaio

Organizações
- Portugal Telecom
- Instituto Superior Técnico
- IST

Lugares
- Castelo Branco
- Serra da Estrela
- Minho

Outros
- Tio Patinhas
- O Lago dos Cisnes
- Renault 4

Qual é o objecto de avaliação?

Sistemas ou módulos de sistemas que façam
Reconhecimento de Entidades Mencionadas (Named Entity Recognition)

Expressões Temporais

Datas
- 24 de Janeiro de 2000
- segundo semestre de 1992
- anos 60

Horas
- meio-dia
- 4 horas da manhã
- 13:40

Expressões Numéricas

Monetárias
- 20 milhões de euros
- 900 mil contos

Percentuais
- 10.5%
- sete por cento

Etapa 1– Anotação Manual ou Semi-Automática

Objectivos

- Identificar as entidades mencionadas que se querem ver reconhecidas pelos sistemas
- Estabelecer um conjunto de etiquetas para classificar as entidades
- Definir critérios para a atribuiçã
- Avaliar o grau de concordância
- Caracterizar um conjunto de pr

Selecção dos textos – CP e CF porque são públicos e estão a ser trabalhados também a outros níveis
(incluir dados sobre as dimensões)

Tarefa
Anotar as entidades mencionadas de forma manual ou semi-automática (automática com revisão) nos 10 primeiros extractos do CETEMPúblico e nos 20 primeiros extractos do CETENFolha

Foi feita a sugestão inicial de usar as etiquetas PESSOA, ORGANIZAÇÃO, LUGAR e OUTRO, deixando em aberto a possibilidade de escolha de um conjunto de etiquetas alternativo ou complementar.

Etapa 1– Anotação Manual ou Semi-Automática

Exemplo

<s>O caso ocorreu numa noite de 1978, na ilha de Carvalo, ao largo da Córsega.</s>
<s>O príncipe jantava com amigos num restaurante deste paraíso para milionários, quando um grupo barulhento de jovens da alta sociedade italiana acostou na enseada de Palma, ao lado do seu iate, o L'Aniram.</s>
<s>Os advogados da defesa sublinharam no processo que este facto perturbou altamente o "senhor de Sabóia".</s>
<s>Naquele ano, as Brigadas Vermelhas (BR) estavam no auge da actividade terrorista, o líder cristão-democrata Aldo Moro acabara de ser raptado, e o príncipe -- proibido de entrar em Itália desde o exílio do pai em 1946 -- teria mesmo recebido ameaças das BR.</s>

Etapa 1– Anotação Manual ou Semi-Automática

Exemplo

<s>O caso ocorreu numa noite de 1978, na ilha de **Carvalo**, ao largo da **Córsega**.</s>
<s>O príncipe jantava com amigos num restaurante deste paraíso para milionários, quando um grupo barulhento de jovens da alta sociedade italiana acostou na enseada de **Palma**, ao lado do seu iate, o **L'Aniram**.</s>
<s>Os advogados da defesa sublinharam no processo que este facto perturbou altamente o "senhor de **Sabóia**".</s>
<s>Naquele ano, as **Brigadas Vermelhas** (**BR**) estavam no auge da actividade terrorista, o líder cristão-democrata **Aldo Moro** acabara de ser raptado, e o príncipe -- proibido de entrar em **Itália** desde o exílio do pai em 1946 -- teria mesmo recebido ameaças das **BR**.</s>

Etapa 1 – Anotação Normativa (Semi-Automática)

Exemplo

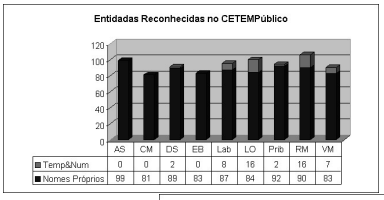
<s>O caso ocorreu numa noite de 1978, na ilha de <NOMEPROP TIPO="LUGAR">**Carvalo**</NOMEPROP>, ao largo da <NOMEPROP TIPO="LUGAR">**Córsega**</NOMEPROP>.</s>
<s>O príncipe jantava com amigos num restaurante deste paraíso para milionários, quando um grupo barulhento de jovens da alta sociedade italiana acostou na enseada de <NOMEPROP TIPO="LUGAR">**Palma**</NOMEPROP>, ao lado do seu iate, o <NOMEPROP TIPO="BARCO">**L'Aniram**</NOMEPROP>.</s>
<s>Os advogados da defesa sublinharam no processo que este facto perturbou altamente o "senhor de <NOMEPROP TIPO="LUGAR">**Sabóia**</NOMEPROP>".</s>
<s>Naquele ano, as <NOMEPROP TIPO="ORGANIZAÇÃO">**Brigadas Vermelhas**</NOMEPROP> (<NOMEPROP TIPO="ORGANIZAÇÃO">**BR**</NOMEPROP>) estavam no auge da actividade terrorista, o líder cristão-democrata <NOMEPROP TIPO="PESSOA">**Aldo Moro**</NOMEPROP> acabara de ser raptado, e o príncipe -- proibido de entrar em <NOMEPROP TIPO="LUGAR">**Itália**</NOMEPROP> desde o exílio do pai em 1946 -- teria mesmo recebido ameaças das <NOMEPROP TIPO="ORGANIZAÇÃO">**BR**</NOMEPROP>.</s>

Etapa 1 – Anotação Normativa (Semi-Automática)

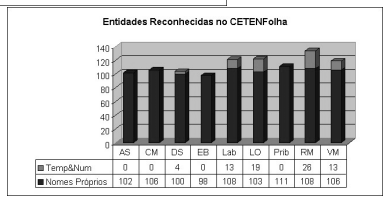
Participantes

- Alberto Simões (Linguateca, Pólo do Minho)
- Cláudia Pinto (Priberam)
- Cristina Mota (Linguateca, Pólo do LabEL)
- Diana Santos (Linguateca, Pólo do Sintef)
- Eckhard Bick (Southern Denmark University)
- Lucelia de Oliveira (NILC)
- Paula Carvalho (LabEL)
- Raquel Marchi (NILC)
- Vanessa Maquiafavel (NILC)

Etapa 1 – Primeiros Resultados



CP	EM	NPr
Mínimo	81	81
Máximo	106	99



CF	EM	NPr
Mínimo	98	98
Máximo	134	111

Etapa 1 – Primeiros Resultados

A propósito, no **Museu da Segunda Guerra Mundial**, que aí foi aberto, a história da maior guerra no continente europeu começa com a fotografia de **Estaline** a cumprimentar o **ministro dos Negócios Estrangeiros** da **Alemanha** nazi, ou seja, a guerra começa com a assinatura do **Pacto Molotov-Ribbentrop**.

A propósito, no **Museu da Segunda Guerra Mundial**, que aí foi aberto, a história da maior guerra no continente europeu começa com a fotografia de **Estaline** a cumprimentar o **ministro dos Negócios Estrangeiros da Alemanha** nazi, ou seja, a guerra começa com a assinatura do Pacto Molotov-Ribbentrop.

Concordância-1
Relativa ao total de **entidades identificadas por pelo menos um** anotador

Concordância-2
Relativa ao total de **nomes próprios identificadas por pelo menos um** anotador

Concordância-3
Relativa ao total de **nomes próprios identificadas por todos** os anotadores

50% (1em 2)

- × Museu da Segunda Guerra Mundial
- ✓Estaline
- • ministro dos Negócios Estrangeiros
- • Alemanha
- × Pacto Molotov-Ribbentrop

- × Museu da Segunda Guerra Mundial
- ✓Estaline
- • ministro dos Negócios Estrangeiros da Alemanha

Etapa 1 – Primeiros Resultados

Concordância entre Anotadores

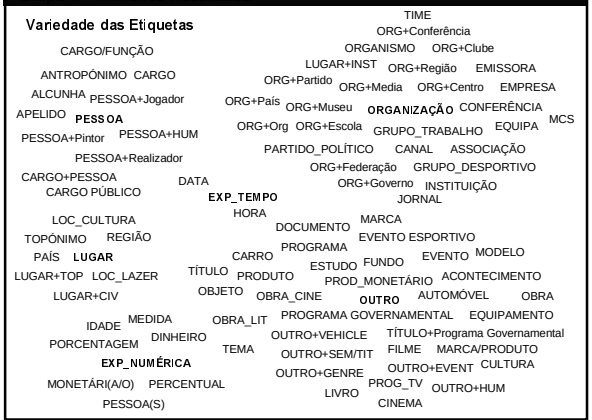
	CETEMPúblico	CETENFolha
Concordância-1	31,58% (30 em 95)	24,59% (30 em 122)
Concordância-2	37,97% (30 em 79)	30,61% (30 em 98)
Concordância-3	50,85% (30 em 59)	45,45% (30 em 66)

Variedade das Etiquetas

	CETEMPúblico	CETENFolha
AS	4	4
CM	6	7
DS	12 + 6 [1]	19 + 9 [1]
EB	4 (11)	4 (11)
Lab	17 [3]	20 [4]
LO	12 [5]	12 [6]
Prib	6 [1]	6
RM	13 [3]	16 [4]
VM	10 [2]	18 [4]

Etapa 1 – Primeiros Resultados

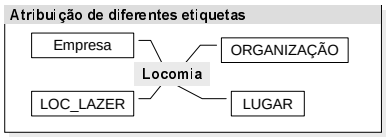
Variedade das Etiquetas



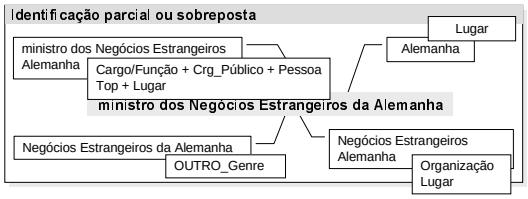
Etapa 1 – Primeiros Resultados

Casos de Discordância

Atribuição de diferentes etiquetas



Identificação parcial ou sobreposta



Etapa 1 – Primeiros Resultados

Casos de Discordância

Ausência de Anotação

Executivo / executivo
Congresso / congresso
Governo / governo
Presidente da República
Ocidente
Estado
Rendimento Mínimo Garantido
secretário de Estado
Primeiro Mundo

Decisões a Tomar

Que sequências considerar como entidades mencionadas?

Nomes próprios? Expressões temporais? Expressões numéricas?

Delimitação de Encaixados?

[Escola de Medicina de Harvard] [Escola de Medicina de [Harvard]]

Incluir cargos, títulos, funções?

Presidente [Jorge Sampaio] [Presidente Jorge Sampaio]

major [Carlos Barbosa] [major Carlos Barbosa]

Atribui-se a etiqueta em função do contexto?

feira especializada que teve lugar em Basileia (**Suíça**)

≠

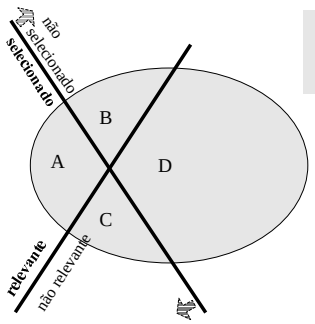
chegará o dia em que a **Rússia** ajudará

Decisões a Tomar

O que fazer quando não é possível decidir?

Anotar / Ignorar Tanto nos recursos de avaliação como nos resultados

Métricas

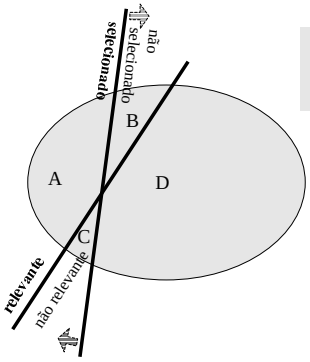


Precision = A / (A+C)

Recall = A / (A+B)

[Rachel Aires, Avaliação De Sistemas De Recuperação De Informação (RI). Panorâmica e Reflexões, 2002]

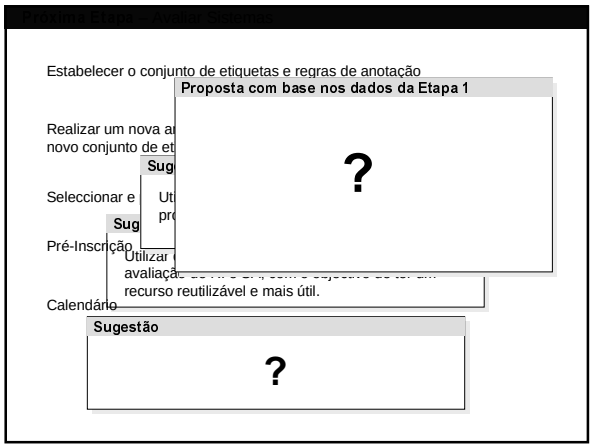
Métricas



Precision = A / (A+C)

Recall = A / (A+B)

[Rachel Aires, Avaliação De Sistemas De Recuperação De Informação (RI). Panorâmica e Reflexões, 2002]


Próxima Etapa – Avaliar Sistema
Estabelecer o conjunto de etiquetas e regras de anotação
Realizar um nova a
novo conjunto de et
Seleccionar e
Pré-Inscrição
Calendário
Proposta com base nos dados da Etapa 1
?
Sug
Sug
recurso reutilizável e mais útil.
Sugestão
?