O impacto de diferentes fontes de conhecimento na marcação de Nomes Próprios em Português

João Tomé da Silva Laranjinho and Irene Pimenta Rodrigues

joao.laranjinho@gmail.com, ipr@di.uevora.pt
Departamento de Informática
Universidade de Évora, Évora, Portugal

Resumo Neste artigo apresenta-se um sistema, independente do domínio, para marcação de nomes de entidades para o português. Este sistema é avaliado de forma a estudar o impacto de diferentes fontes de conhecimento nos resultados do sistema.
O marcador usa informação morfo-sintáctica, sintáctica e semântica. A informação morfo-sintáctica vem de um dicionário local que completa a sua informação recorrendo a dicionários disponíveis na rede como o da Priberam. A informação sintáctica das frases vem de uma gramática construída para analisar frases interrogativas, esta gramática tem bons resultados para as frases interrogativas (acima de 95% de cobertura) mas para as frases dos outros corpura testados tem um mau desempenho (abaixo dos 30% de cobertura). A informação semântica usada nas experiências de avaliação vem da Wikipédia.
Na avaliação do sistema e do impacto das diferentes fontes de informação usaram-se três corpura distintos: um conjunto de documentos com 700 perguntas do CLEF; 300 frases de noticias do Público; e 200 frases do corpus CD do segundo Harem.

Abstract We present a domain independent system that identifies proper names in Portuguese texts. This system is evaluated in order to study the impact of the different knowledge sources on its performance. The knowledge sources are: morph-syntactic obtained from a local dictionary that is able to consult online dictionaries; syntactic from a Portuguese grammar that indicates if a sentence with some proper names tagged is correct; and semantic obtained from an online encyclopedia, the Wikipedia.
In the evaluation, we use three different corpus: a set of documents with 700 questions from CLEF, 300 sentences of news from Publico, and 200 sentences from the corpus of the second Harem. The Portuguese grammar was developed in order to analyse Portuguese interrogative sentences. This fact is reflected in the performance of our system in the different corpus since the coverage of the grammar is up to 95% for CLEF and bellow 30% for the others.

1 Introdução

Os sistemas de marcação de nomes de entidades são um caso particular (a identificação) dos sistemas de reconhecimento de entidades mencionadas (REM) que identificam e classificam as entidades de acordo com uma hierarquia pré-definida de categorias. Para

INForum 2010 - II Simpósio de Informática, Luís S. Barbosa, Miguel P. Correia (eds), 9-10 Setembro, 2010, pp. 561–572

o inglês, alguns dos sistemas actuais conseguem um valor de 93% na medida F em foruns de avaliação como o MUC-7 [NS07]. Para o português, no segundo Harem, o melhor sistema, o sistema da Priberam [AFM+08] consegue atingir cerca de 71% para a medida F na identificação de entidades mencionadas (ou marcação de nomes de entidades).

O sistema REMUE (Reconhecimento de Entidades Mencionadas da Universidade de Évora) foi, inicialmente, desenvolvido para melhorar o desempenho de um sistema de resposta automática a perguntas [QRPV06], em particular para auxiliar na escolha das melhores análises sintácticas de uma pergunta em português. A avaliação inicial do REMUE foi feita usando um corpus com as perguntas do CLEF na tarefa de pergunta-resposta [SR04] de várias edições. A marcação dos nomes de entidades neste corpus foi feita manualmente para a avaliação do REMUE. Para estudar o comportamento do REMUE noutro tipo de corpus construiu-se uma amostra com frases do Público que também foram anotadas manualmente. E finalmente fez-se uma avaliação com uma amostra da colecção dourada do Segundo Harem que apesar de não permitir uma comparação directa com os sistemas que concorreram ao Harem nos permite ver o impacto das diferentes fontes de conhecimento nos resultados do REMUE (o REMUE não foi avaliado no Segundo Harem).

O REMUE usa três tipos de fontes de conhecimento para decidir a marcação de nomes de entidades:

- Informação morfo-sintáctica - num dicionário local que completa a sua informação recorrendo a dicionários disponíveis na rede como o Dicionário da Priberam.
- Informação sintáctica - o resultado de um analisador sintáctico, a estrutura sintáctica das frases com os nomes de entidades das frases. Uma frase pode ter zero, uma ou mais estruturas sintácticas associadas. A estrutura sintáctica pode ser parcial ou total. Esta informação é usada, pelo REMUE, para decidir a melhor marcação de nomes de entidades na frase.
- Informação semântica - informação de dicionários e enciclopédias que indicam se o nome da entidade existe nalgum contexto. No REMUE por enquanto só se usa a informação da Wikipédia que como se pode ver na sua avaliação tem um grande impacto na correcção das marcações.

As técnicas usadas nos sistemas de REM dividem-se em: modelos baseados em regras e modelos estatísticos. Actualmente a técnica dominante é a baseada em modelos estatísticos e aprendizagem. Estes sistemas requerem grandes quantidades de dados anotados manualmente para a fase de treino e alguns sistemas tem um desempenho dependente do domínio.

Os modelos baseados em regras têm apresentado melhores resultados (ver MUC-7), no entanto, requerem muito trabalho feito por linguistas na sua implementação e muitas vezes o seu desempenho depende do domínio.

Algum dos recursos utilizados no REM [AA08] são:

- Corpus - conjuntos de textos anotados. Normalmente em conjunto com os corpura, são utilizadas estratégias de aprendizagem, como por exemplo: modelos de Markov não observáveis (Hidden Markov Models - HMM), árvores de decisão, modelos de

máxima entropia, SVMs [NS07]. Um dos problemas da construção de um corpus está relacionado com a anotação;
- Almanaques - dicionários de entidades mencionadas (EM);
- Metapalavras - representam as palavras próximas das EM. Estas são palavras que dão alguma informação sobre as entidades, normalmente são utilizadas na fase de desambiguação. Exemplos: escritor, rua, etc;
- Abreviaturas - palavras que fazem parte das entidades e dão informação sobre a classe da entidade. Normalmente são utilizadas na classificação. Exemplo: Dr., Sr., etc;
- Regras de similaridade - um conjunto de regras que definem semelhanças entre as entidades a serem classificadas e as entidades que existem em almanaques.

O desempenho de um sistema de REM pode ser medido com diversas métricas [SC07] que representam o desempenho em valores numéricos.

As três métricas que normalmente são utilizadas para avaliar o desempenho de um sistema de recolha de informação são as seguintes: abrangência (Recall), precisão (Precision) e medida F (F-Measure).

- A abrangência mede a relação entre o número de resultados correctos e o número de resultados existentes. A fórmula da Abrangência é a seguinte:
  Abrangência = $\frac{Resultados\ Correctos\ \cap\ Resultados\ Existentes}{Resultados\ Existentes}$
- A precisão mede a relação entre o número de resultados correctos e o número de resultados obtidos. A fórmula da Precisão é a seguinte:
  Presisão = $\frac{Resultados\ Correctos\ \cap\ Resultados\ Obtidos}{Resultados\ Obtidos}$
- A medida F é uma média harmónica entre a precisão (P) e a abrangência (A). A fórmula da *medida F* é a seguinte:
  Medida-F = $2 * \frac{P * A}{P + A}$

Na próxima secção, 2, apresenta-se a arquitectura do REMUE com os seus módulos de processamento e as diferentes fontes de conhecimento. Na secção 3, apresentam-se os testes feitos com o REMUE em 3 corpura diferentes procurando ver o impacto das diferentes fontes de conhecimento em cada um dos corpura. Para estudar o impacto fazem-se 8 testes diferentes calculando a precisão, a cobertura e a medida F para cada corpus. Finalmente, na secção 4, analisam-se os resultados da avaliação feita e discutem-se alguns aspectos que podem ser melhorados no REMUE e na sua avaliação.

2 Arquitectura do sistema do REMUE

O REMUE recebe um ficheiro de texto para marcação de nomes de entidades e retorna dois ficheiros, um com o texto em que os nomes de entidades surgem marcados e um outro com a lista de nomes de entidades de cada frase.

Para a marcação dos nomes de entidades recorre-se a regras que codificam as preferências na escolha dos nomes de entidades. As regras usam informação sobre:

- O número de átomos do nome de entidade;

– Informação sobre a primeira letra das palavras (maiúscula ou minúscula);
– A classe morfo-sintáctica de cada palavra;
– Informação sobre alguns caracteres como: aspas, números e sinais de pontuação.

O REMUE contém 4 módulos. Na Figura 1 é apresentada a arquitectura do REMUE.

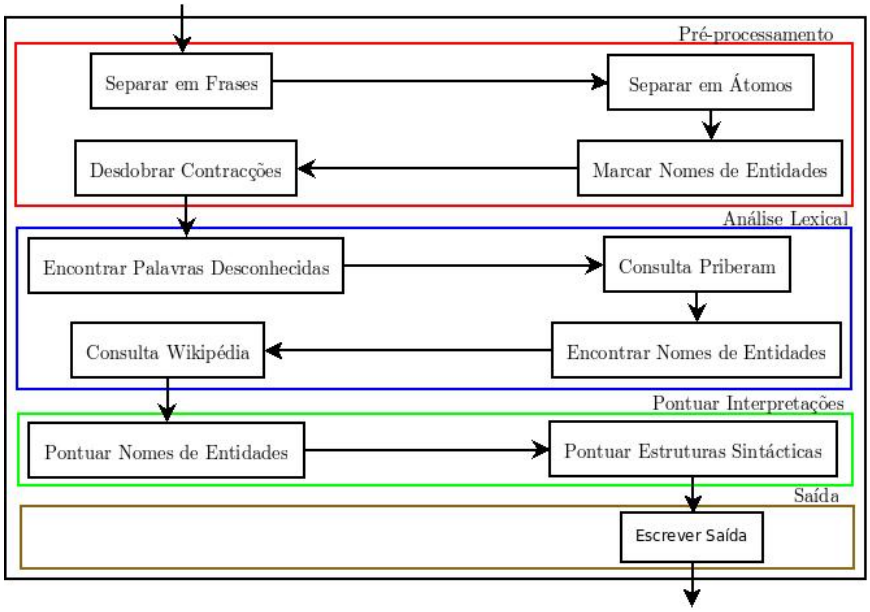

Figura 1. Arquitectura do REMUE

Os módulos são: pré-processamento, análise lexical, pontuar interpretações e saída.

2.1 Pré-processamento

Um texto para ser analisado deve encontrar-se separado em frases. Uma frase é constituída por palavras, números, sinais de pontuação, etc.

Na separação de um texto em frases utiliza-se uma estrutura que guarda, em cada posição, o conjunto de interpretações de uma frase.

Uma interpretação de uma frase para ser analisada deve encontrar-se separada em átomos. Um átomo é constituído por uma sequência de caracteres sem espaços em branco. Os átomos podem ser palavras, números, sinais de pontuação, sequências de letras alternadas com números, etc.

Um candidato a nome de entidade é constituído por uma ou várias palavras que começam com maiúscula, podendo possuir números e elementos de ligação de nome de

entidades. Os elementos de ligação de nome de entidade considerados são os seguinte: “de”, “da”, “do”, “das”, “dos”, “e”, “&” e “-”.

O sistema para uma interpretação que tem N candidatos a nomes de entidades produz 2^N interpretações. As diferenças entre às interpretações são ao nível do número de candidatos a nomes de entidades marcados que variam entre 0 e N.

2.2 Análise Lexical

Na análise lexical reúne-se num ficheiro as palavras que não se encontram no dicionário local.

O dicionário local contém informação (entradas) sobre palavras e nomes de entidades, sendo a informação sobre as palavras relativa às classes gramaticais, além de valores para variáveis morfo-sintáctico-semânticas. Enquanto que a informação sobre os nomes de entidades é relativa a sua existência.

As entradas no dicionário local são as seguintes: adjectivo (adj), advérbio (adv), determinante (det), substantivo comum (nome), número (num), preposição (prep), contracção (contr), conjunção (conj), interjeição (interj), pronome (pron), verbo (verbo) e nome de entidade (nomeEntidade).

O Priberam[1] é um dicionário disponível na rede, no qual podem ser feitas consultas sobre palavras através da Internet. Em cada consulta obtém-se informações morfo-sintáctico-semântica acerca da palavra pesquisada.

Para consultar, o Priberam foi desenvolvida uma aplicação que estabelece a ligação ao mesmo. A aplicação recebe o ficheiro de texto com as palavras. Para cada palavra estabelece-se uma ligação ao Priberam obtendo-se um ficheiro com a informação da palavra.

A informação da palavra no Priberam consiste numa tabela. Cada tabela contém duas partes (2 tr's). Na primeira parte encontra-se informação directa sobre a palavra em pesquisa. Na segunda parte além de existir informação directa, existe também informação indirecta. A informação indirecta é complementada com ligações para outras palavras que se encontram relacionadas com a palavra em pesquisa.

Por exemplo, informação que se encontra no Priberam sobre a palavra “assassino”, é a que se encontra na Tabela 1.

Relativamente à tabela com a informação da palavra “assassino”, na primeira parte da tabela é refido que esta palavra é um ajectivo masculino e um substantivo comum de género masculino. Na segunda parte é referido que a palavra “assassino” é um verbo que se encontra na 1ª pessoa do singular do presente do indicativo. Ainda na segunda parte da tabela é feita referência ao verbo “assassinar” que é o infinitivo da conjugação “assassino”.

Para processar a informação do Priberam foi criado um analisador sintáctico. O analisar sintáctico por cada pesquisa cria uma estrutura com o nome da palavra pesquisada e uma lista com as características da mesma.

As estruturas obtidas no processamento da informação de Priberam devem ser convertidas em entradas para o dicionário local, podendo uma estrutura dar origem a uma ou mais entradas.

[1] http://priberam.pt/dlpo/dlpo.aspx

```
<xml search="assassino">
  <table style="background-color:\#eee; width:100%;"
    cellpadding="4"
    cellspacing="0"
    border="0"
    bordercolor="\#cccccc">
    <tr>
      <td>
        <div>
          <b>assassino</b> |<em>adj.</em> |<em>s. m.</em>
        </div>
      </td>
    </tr>
    <tr>
      <td>
        <div>1ª pess. sing. pres. ind. de
          <a href="&#xA; default.aspx?pal=assassinar">assassinar</a>
        </div>
      </td>
    </tr>
  </table>
</xml>
```

Tabela 1. Informação da palavra "assassino" no dicionário da Priberam

A Wikipédia[2] é uma enciclopédia, na qual podem ser feitas consultas via Internet sobre nomes de entidades e outras palavras. Esta enciclopédia pode ser utilizada para verificar se determinado nome de entidade existe.

A informação sobre a existência de uma entrada para um nome de entidade na Wikipédia é feita em 2 passos.

No primeiro passo é estabelecida uma ligação à Wikipédia, que verifica se existem entradas para o nome da entidade, devolvendo-se um valor Booleano, "true" ou "false", consoante tenha ou não entrada na Wikipédia.

No segundo passo gera-se a entrada para o dicionário local, no caso de existir entrada do nome da entidade na Wikipédia (devolvido "true").

2.3 Heurística para Pontuar Interpretações

A marcação dos candidatos a nomes de entidades numa frase pode produzir várias interpretações. A melhor interpretação de uma frase é aquela que contém o maior número de nomes de entidades marcados correctamente e que pode ser representada por uma estrutura sintáctica (interpretação com análise sintáctica). Para encontrar a melhor interpretação de uma frase pode recorrer-se a uma função heurística que pontua cada interpretação.

[2] http://pt.wikipedia.org/wiki/Página_principal

Nos candidatos a nomes de entidades existem características comuns. Essas características podem ser utilizadas para agrupar os candidatos com características iguais e atribuir-lhes iguais pontuações.

Os candidatos a nomes de entidades foram divididos em:

- NE WIKI - candidato a nome de entidade que tem entrada na Wikipédia;
- NE SIMPLES - candidato a nome de entidade que não tem entrada na Wikipédia e, é composto por apenas uma palavra das seguintes classes gramaticais: substantivo comum, adjectivo ou nome próprio;
- NE COMPOSTO - candidato a nome de entidade que não tem entrada na Wikipédia e, é composto por mais que um átomo, em que o primeiro átomo é uma palavra que pertence a uma das seguintes classes gramaticais: substantivos comum, adjectivos, verbo ou nome próprio;
- NE NUM - candidato a nome de entidade que não tem entrada na Wikipédia e, é composto por um valor numérico;
- NE ASPAS - candidato a nome de entidade delimitado por aspas (“”);
- NE DATA - candidato a nome de entidade marcado como data;
- NE HORA - candidato a nome de entidade marcado como hora;
- NAO NE - candidato a nome de entidade que não reúne nenhuma das características anteriores.

As interpretações de frases que têm análise sintáctica, ou seja, uma ou mais estruturas sintácticas, devem ser valorizadas face às que não têm. A pontuação atribuída a essas interpretações, é um passo importante na escolha da melhor interpretação.

As interpretações que têm uma ou mais estruturas sintácticas foram divididas em:

- TOTAL REP - interpretação totalmente representável por uma estrutura sintáctica;
- PARCIAL REP - interpretação parcialmente representável por uma estrutura sintáctica;

O REMUE recebe um conjunto de interpretações de cada frase, pontuando os candidatos a nomes de entidades de cada interpretação e a interpretação em termos de análise sintáctica.

Para a análise sintáctica desenvolveu-se um analisador sintáctico, recorrendo às gramáticas de cláusulas definidas (Definite Clause Grammars - DCGs).

Este analisador sintáctico verifica se as interpretações das frases de um texto são representadas por estruturas sintácticas e infere essas estruturas.

Na construção da estrutura, por cada palavra da frase a analisar são verificadas as suas características no dicionário local.

2.4 Saída

A saída de um ficheiro processado pelo REMUE é constituída pela interpretação mais correcta de cada frase, ou seja, a interpretação que recebeu maior pontuação da função heurística.

Na saída, são gerados dois ficheiro, um com o texto em que os nomes de entidades são destacados das restantes átomos com etiquetas e um outro ficheiro que contém por linha: o número da frase, a lista de nomes de entidades e o valor de heurística atribuído.

3 Avaliação

Na avaliação do REMUE utilizaram-se 3 métricas que são usadas na avaliação de sistema de recolha de informação: precisão, cobertura e medida F. Estas métricas foram adaptadas ao problema de marcação de nomes de entidades.

Na avaliação comparam-se os ficheiros com a lista de nomes de entidades marcados manualmente e a lista de nomes de entidades marcados pelo REMUE.

O resultado da avaliação é escrito, frase a frase, num ficheiro com as seguintes variáveis: número da frase, número de nomes de entidades marcados, número de nomes de entidades correctos dos marcados, número de nomes de entidades incorrectos dos marcados, número de nomes de entidades existentes, precisão, cobertura, medida F e valor da heurística.

Na avaliação do REMUE foram utilizados 3 corpura: 700 perguntas do CLEF, 300 notícias do jornal o Público e 200 frases do Harem.

O CLEF (Cross-Language Evaluation Forun) é uma série de avaliações conjuntas que promove a pesquisa e desenvolvimento na área de recolha de informação entre várias línguas. A participação do português tem sido financiada pela Linguateca, a nível de recursos humanos, e pelo diário Público (Portugal) e Folhas de São Paulo (Brasil), a nível de fornecimento de recursos. A Linguateca disponibiliza a colecção CHAVE que contém textos, tópicos e perguntas utilizados nas edições do CLEF.

O corpus do CLEF foi usado para verificar até que ponto o uso da gramática desenvolvida para frases interrogativas tem impacto no desempenho do REMUE neste corpus.

O corpus do CLEF foi anotado manualmente identificando os nomes de entidades das frases, ou seja, sem classificação.

O corpus do Público foi utilizado para testar o desempenho do REMUE em frases para as quais a gramática tem um mau resultado, não consegue obter análise sintáctica em mais de 70% das frases do corpus.

O corpus do Segundo Harem foi utilizado para obter uma avaliação independente, ou seja, com as anotações feitas pela equipe da Linguateca. Apesar de usarmos as marcações do Harem, a avaliação foi feita por nós e só para uma amostra da Colecção Dourada do Segundo Harem, o que nos impede de comparar os nossos resultados com os dos sistemas que concorreram ao Segundo Harem. Fizemos os testes sobre uma amostra do Segundo Harem porque a nossa gramática que foi construída para frases interrogativas tem um mau desempenho (tempo e espaço) na análise das frases deste corpus.

Com cada um dos conjuntos de textos foram realizados 8 testes, nos quais foram feitas variações das pontuações atribuídas aos grupos de candidatos a nomes e às interpretações de frases representáveis por estruturas sintácticas.

Com estes testes procuramos estudar o impacto da informação: morfo-sintáctica, semântica e sintáctica no desempenho do sistema.

3.1 Heurística Utilizada: Pontuações Atribuídas

Em baixo pode ver-se a heurística usada no Teste 1. Os valores da heurística correspondem aos que dão o melhor desempenho ao REMUE nos diferentes corpura. Estes valores foram encontrados por tentativa e erro e não garantimos que sejam os melhores.

- NE WIKI $= 10 * (1 + Es)$ (Es -número de átomos).
  Neste teste valorizamos o facto de os nomes de entidades terem entrada na Wikipédia e também preferimos os nomes de entidades mais compridos (com mais átomos) se existirem na Wikipédia.
- NE SIMPLES $= 4$ e NE COMPOSTO $= 9 * (1 + Es)$ (Es -número de átomos)
  Valorizamos o comprimento dos nomes de entidades mesmo quando não temos entrada na Wikipédia.
- NE NUM$= -9$
  Desvalorizamos números isolados que sejam marcados como nomes de entidades.
- NE ASPAS $= 15$, NE DATA $= 15$ e NE HORA $= 15$
  Valorizamos expressões entre aspas que sejam marcadas como nomes de entidades e outras expressões que sejam marcadas como data e hora.
- NAO NE $= -10 * (1 + Es)$ (Es -é número de espaços em branco numa cadeia de caracteres)
  Desvalorizamos expressões que não satisfazem os nomes de entidades anteriores.
- TOTAL REP $= 100$ e PARCIAL REP $= 60$
  Com estes valores, valorizamos bastante o facto de a frase com os nomes das entidades marcados ter análise sintáctica (TOTAL REP=100 e PARCIAL REP=60)

Na Tabela abaixo podem ver-se os parâmetros usados na definição da heuristíca para cada teste.

	Heuristica
NE WIKI	NE1*(1+Es)
NE SIMPLES	NE2
NE COMPOSTO	NE3*(1+Es)
NE NUM	NE4
NE ASPAS	NE5
NE DATA	NE6
NE HORA	NE7
NAO NE	NE8*(1+Es)
TOTAL REP	I1
PARCIAL REP	I2

	Teste1	Teste2	Teste3	Teste4	Teste5	Teste6	Teste7	Teste8
I1	100	0	100	100	100	100	100	100
I2	60	0	60	60	60	60	60	60
NP1	10	10	0	10	10	10	10	10
NP2	4	4	4	5	4	4	4	4
NP3	9	9	9	5	9	9	9	9
NP4	-9	-9	-9	-9	0	-9	-9	-9
NP5	15	15	15	15	15	0	15	15
NP7	15	15	15	15	15	15	0	15
NP8	15	15	15	15	15	15	15	0
Es	N	N	N	0	N	N	N	N

Com estes testes pretendemos ver o impacto:

- teste 2 – da análise sintáctica. Não se tem em conta a existência de estrutura sintáctica para a frase com os nomes de entidades.
- teste 3 – da Wikipédia. Não se tem em conta a informação sobre a existência de entrada para o nome da entidade na Wikipédia.
- teste 4 – do comprimento do nome da entidade. Não se valoriza o número de átomos do nome da entidade.
- teste 5 – dos números isolados. Não se valorizam os números isolados que sejam marcados como nomes de entidades.
- teste 6 – das expressões entre aspas. Não se valorizam as expressões entre aspas que sejam marcadas como nomes de entidades.
- teste 7 – das datas. Não se valorizam as expressões que sejam marcadas como datas.
- teste 8 – das horas. Não se valorizam as expressões que sejam marcadas como horas.

	Prec	Cob	Med-F
Teste 1	0,9836	0,9804	0,9808
Teste 2	0,9765	0,9746	0,9727
Teste 3	0,9315	0,9335	0,9313
Teste 4	0,8949	0,9029	0,8975
Teste 5	0,9779	0,9796	0,9780
Teste 6	0,9746	0,9768	0,9748
Teste 7	0,9779	0,9796	0,9780
Teste 8	0,9779	0,9796	0,9780

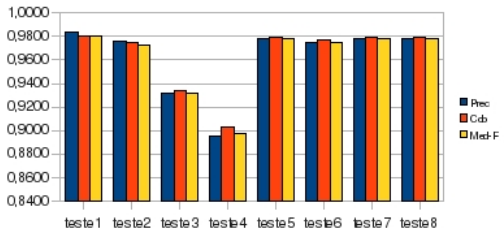

Figura 2. Testes com o corpus do CLEF

4 Conclusões e Trabalho Futuro

Na avaliação do REMUE foram utilizados 3 corpura: 700 perguntas do CLEF, 300 notícias do jornal o Público e 200 frases do Harem. Com cada um destes corpura foram realizados 8 testes com variações de alguns dos parâmetros que pontuam os nomes de entidades marcados e as interpretações de frases com análise sintáctica.

Como se pode ver na secção 3, onde se apresentam os resultados da avaliação do REMEU para 3 domínios diferentes de frases, o teste 1 foi aquele que apresentou os

	Prec	Cob	Med-F
Teste 1	0,9461	0,9547	0,9457
Teste 2	0,9450	0,9547	0,9450
Teste 3	0,8931	0,9238	0,8989
Teste 4	0,6945	0,7750	0,7210
Teste 5	0,8743	0,9547	0,8975
Teste 6	0,9003	0,9259	0,9042
Teste 7	0,9304	0,9547	0,9337
Teste 8	0,9304	0,9547	0,9337

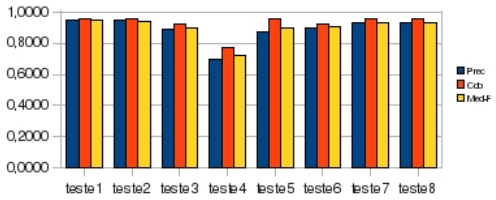

Figura 3. Testes com corpus do Público

	Prec	Cob	Med-F
Teste 1	0,7659	0,8877	0,8223
Teste 2	0,7559	0,8877	0,8165
Teste 3	0,6684	0,7194	0,6929
Teste 4	0,7126	0,8397	0,7710
Teste 5	0,7630	0,8827	0,8185
Teste 6	0,7582	0,8757	0,8127
Teste 7	0,7659	0,8827	0,8202
Teste 8	0,7659	0,8827	0,8202

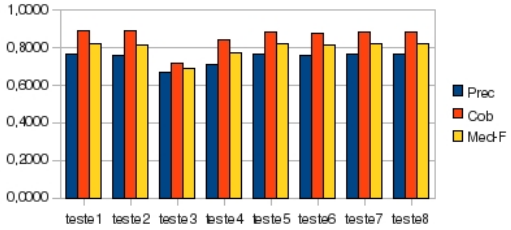

Figura 4. Testes com corpus do HAREM

melhores resultados nos 3 corpura. Neste teste valorizamos os nomes de entidades que se encontram na Wikipédia, as expressões entre aspas, as datas e as horas. Além disso, também valorizamos as interpretações de frases que contêm estruturas sintácticas.

No teste 2 não temos em contam se com os nomes de entidades marcados, a frase tem análise sintáctica. Só nos corpura do CLEF e Harem se vê diferença, 0.8% e 1.0%, entre o teste 1 e o teste 2, no corpus do Público a diferença é quase nula. Como a gramática utilizada não tem um bom desempenho nas frases do Público e do Harem, vamos procurar repetir os testes com uma gramática que tenha um melhor desempenho para estes textos. No entanto podemos concluir que o uso de uma gramática pode melhorar o desempenho do REMUE, ainda que de forma ligeira. Também podemos concluir que o uso da gramática nunca baixa o desempenho do REMUE.

O teste 3 também demonstra que nos 3 corpura os resultados baixam, 16% na cobertura do Harem, 3% na cobertura do Público e 4% na cobertura do CLEF. Permite-nos concluir que o uso de uma enciclopédia melhora os resultados.

O teste 4 retira o peso ao comprimento dos nomes de entidades, os resultados nos diferentes corpura permite concluir que escolher os nomes mais compridos é uma boa estratégia.

Os outros testes mostram que esta informação não tem grande impacto no desempenho do REMUE.

O REMUE pode evoluir em diferentes direcções mas as mais imediatas incluem: utilizar uma gramática com maior cobertura, transportar o REMUE para o inglês, e incluir a classificação das entidades mencionadas.

Referências

[AA08] Marcelo Adriano Amancio and Sandra Maria Aluísio. Explicitação de entidades mencionadas visando o aumento da inteligibilidade de textos em português. Technical report, Universidade de São Paulo, Agosto de 2008.

[AFM+08] Carlos Amaral, Helena Figueira, Afonso Mendes, Pedro Mendes, Cláudia Pinto, and Tiago Veiga. Adaptação do sistema de reconhecimento de entidades mencionadas da priberam ao harem. In Cristina Mota and Diana Santos, editors, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca, 2008.

[NS07] David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Linguisticae Investigationes*, 1(30):3–26, 2007.

[QRPV06] Paulo Quaresma, Irene Rodrigues, C. A. Prolo, and R. Viera. Um sistema de pergunta-resposta para uma base de documentos. *Letra de Hoje - Revista da Pontifíca Universidade Católica do Rio Grande do Sul*, 41(2):43–63, Junho 2006.

[SC07] Diana Santos and Nuno Cardoso, editors. *Reconhecimento de entidades mencionadas em português*. Linguateca, 2007.

[SR04] Diana Santos and Paulo Rocha. Chave: topics and questions on the portuguese participation in clef. In C. Peters and F. Borri, editors, *Cross Language Evaluation Forum: Working Notes for the CLEF 2004 Workshop*, pages 639–648, Bath, UK, September 2004 2004.