# Marcação de Nomes Próprios usando técnicas de pesquisa local e recorrendo a fontes de conhecimento na Internet

João Laranjinho
Universidade de Évora
Évora, Portugal
joao.laranjinho@gmail.com

Irene Rodrigues
Universidade de Évora
Évora, Portugal
ipr@di.uevora.pt

Lígia Ferreira
Universidade de Évora
Évora, Portugal
lsf@di.uevora.pt

**Abstract**

Neste artigo apresenta-se um sistema, independente do domínio, para marcação de nomes próprio para o português e inglês. O sistema é avaliado de forma a estudar o impacto de diferentes fontes de conhecimento nos resultados.

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 e do LookWayUP. A informação semântica usada nas experiências de avaliação vem da Wikipédia e do WordNet. No sistema são usadas também algumas das técnicas pesquisa local na marcação de nomes próprios.

Na avaliação do sistema e do impacto das diferentes fontes de informação usaram-se frases 2 corpora: 100 frases do WSJ e 100 frases do Brown usadas na fase de treino e 100 frases do Brown usadas na fase de testes.

## 1 Introdução

O sistema que apresentamos chama-se REMUE2011. Este sistema é uma evolução do REMUE [1] que tinha como objectivo a marcação de nomes próprio para o Português. Actualmente além de marcar nomes próprios para o Português também marca nome próprios para o Inglês.

Na decisão de marcar nomes próprios usam-se 2 tipos de fontes de conhecimentos:

- Informação morfo-sintáctica — do dicionário local que é completada recorrendo a dicionários que estão disponíveis na Web como o dicionário Priberam [1] (para o Português) e o Look Way Up [2] (para o Inglês).
- Informação semântica — de dicionários e enciclopédias como a Wikipédia [3] e o WordNet [4] que indicam se o nome próprio existem em algum contexto.

## 2 Arquitectura do Sistema

A arquitectura do REMUE2011 contém 4 módulos. Na Figura 6 é apresentada a arquitectura com os seus módulos: pré-processamento, análise lexical, pesquisa local e saída.

No pré-processamento separa-se o texto em frases e as frases em átomos. As frases são constituídas por átomos e os átomos por sequências de caracteres.

Na análise lexical consulta-se em dicionários on-line a informação morfo-sintáctico-semântica das palavras que não se encontram no dicionário local, guardando-se essa informação no dicionário local.

José Saias, Luís Rato and Teresa Gonçalves (eds.): JIUE 2011, 2011, volume 1, issue: 1, pp. 1-6

[1] http://www.wikipedia.org/

[2] http://lookwayup.com/free/

[3] http://www.wikipedia.org/

[4] http://wordnetweb.princeton.edu/perl/webwn?s

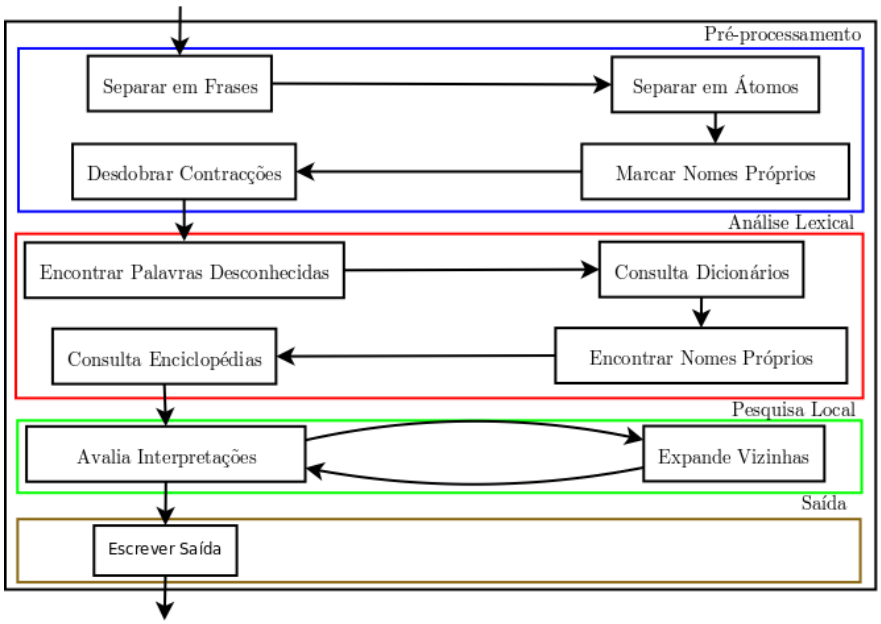



Figure 1: Arquitectura do REMUE

Na pesquisa local gera-se um conjunto de interpretações iniciais e posteriormente avaliam-se estas. Na avaliação quando uma interpretação contém vizinhas com valor de heurística superior, expandem-se as vizinhas e em seguida avaliam-se. A avaliação termina quando não são encontradas mais interpretações vizinhas com valor de heurística superior.

Finalmente na saída utiliza-se a interpretações que obteve o valor mais alto de heurística.

## 3 Função de Avaliação

Na função de avaliação estudou-se o impacto de diferentes fontes de conhecimento na marcação de marcação de nomes próprios.

O estudo das fontes incidiu sobre: o nome próprio, o átomo anterior ao nome próprio e átomo posterior ao nome próprio.

- No nome próprio estudou-se o impacto de:
  - — estar na Wikipédia (WIKI), no WordNet (WORDNET) ou parte do nome na Wikipédia (WIKI_P);
  - — conter maiúsculas (MAIUSCULAS) ou números (NUMEROS);
  - — conter adjectivos(ADJS), advérbios (ADVS), conjunções (CONJS), determinantes (DETS), nomes comuns (NOME), preposições (PREPS), pronomes (PRONS), verbos (VERBOS) ou palavras desconhecidas (DESCS);
  - — o números de átomos (ATOMOS);
- No átomo anterior e posterior ao nome próprio estudou-se o impacto de:
  - — estar na Wikipédia (WIKI) e WordNet (WORDNET);
  - — conter maiúsculas (MAIUSCULAS) ou números (NUMEROS);
  - — ser adjectivo (ADJ), advérbio (ADV), conjunção (CONJ), determinante (DET), nome comum (NOME), preposição (PREP), pronome (PRON), verbo (VERBO) ou uma palavra desconhecida (DESC);

# 4 Marcador Local

Inicialmente o marcador de nomes próprio foi criado com o intuito de analisar todas as interpretações de uma frase. No entanto, em presença de frases com muitas interpretações torna-se impraticável analisar todas as interpretações devido ao elevado tempo de processamento.

Para resolver este problema pensou-se em aplicar técnicas de pesquisa local no marcador de nomes próprios, sendo um estado (interpretação) constituído por um conjunto de tuplo em que cada tuplo consiste de um átomo (palavra, sinal de pontuação, números, etc) e um valor numérico (0 ou 1) que indica se o átomo pertence ou não a um nome próprio, e uma interpretação vizinha uma nova interpretação que é gerada a partir de uma outra através da mudança de um dos valores numéricos (mutação) de um tuplo ou de vários (mutações).

O algoritmo recebe uma interpretações e aleatoriamente atribui valores que indicam se os átomos da interpretação pertencem ou não a nomes próprios.

Para encontrar a melhor interpretação de uma frase utiliza-se o algoritmo *search_best_interpretation*:

```
interpretation function search_best_interpretation(interpretation s,number Max_Flips,table t)
  interpretation b, i, i1
  b ← s
  for j := 1 to Max_Flips do
    i ← random_interpretation(s)
    insert(i, t)
    i1 ← neighbor_interpretation(i,t)
    while i1! = null do
      if h(i1) >= h(i)
        i ← i1
        i1 ← neighbor_interpretation(i1,t)
        insert(i1, t)
      else
        i1 ← neighbor_interpretation(i,t)
        insert(i1, t)
      endif
    endwhile
    if h(i) >= h(b)
      b ← i
    endif
  endfor
  return b
end search_best_interpretation
```

O algoritmo *search_best_interpretation* recebe uma frase sem nomes próprios marcados, o número de interpretações iniciais que deve gerar aleatoriamente e uma tabela onde guarda as interpretações que gera. Para cada interpretação inicial são expandidas as interpretações vizinhas com valores heurística superior e por sua vez as vizinhas das vizinhas. Finalmente quando todas estas interpretações forem analisadas é retornada a interpretação que obteve o valor mais alto de heurística.

# 5 Optimizador Local

Para não se apurar manualmente os valores dos parâmetros da função heurística consoante o corpora que se pretende analisar, criou-se uma aplicação que automaticamente encontra valores para os parâmetros que maximizam o desempenho da função heurística.

Inicialmente a aplicação foi criada com o intuito de estudar todos os conjuntos de parâmetros. No entanto, verificou-se que aumentando o número de parâmetros não se conseguia estudar todos esses conjuntos devido ao tempo de processamento.

Para contornar esse problema pensou-se em incorporar na aplicação técnicas de pesquisa local, sendo um estado constituído por um conjunto de parâmetros em que cada parâmetro é representado por um valor numérico e um estado vizinho um novo conjunto de parâmetros que contém uma diferença (mutação) ou várias (mutações) em relação ao estado anterior.

Para apurar os parâmetros utiliza-se o algoritmo *determine_parameters*:

```
parameters function determine_parameters(parameters s,number Max_Flips, table t)
  parameters b, ps, ps1
  b ← s
  for j := 1 to Max_Flips do
    ps ← random_parameters(s)
    insert(ps,t)
    ps1 ← neighbor_parameters(ps,t)
    while ps1 != null do
      insert(ps1,t)
      if ht(ps1) >= h(ps)
        ps ← ps1
        ps1 ← neighbor_parameters(ps1,t)
      else
        ps1 ← neighbor_parameters(ps1,t)
      endif
    endwhile
    if ht(ps) >= ht(b)
      b ← ps
    endif
  endfor
  return b
end determine_parameters
```

O algoritmo *determine_parameters* recebe um conjunto de parâmetros vazio, o número de conjuntos iniciais de parâmetros que pode gerar aleatoriamente e uma tabela onde guarda os conjuntos de parâmetros que gera. Para cada conjunto de parâmetros inicial são expandidos os conjuntos de parâmetros vizinhos que dão valor de heurística superior ao texto e por sua vez os vizinhos dos vizinhos. Finalmente quando todos estes conjuntos de parâmetros forem analisados é retornado o conjunto que deu o valor mais alto de heurística ao texto.

## 6 Avaliação

Na avaliação inicialmente estudou-se o impacto de isolar cada uma das fontes de conhecimento. Na seguinte tabela podem ver-se os valores alcançados nas métricas de precisão, cobertura e medida-F quando se optimizaram os parâmetros usando no treino 70 frases com 4 nomes próprios no máximo por frase retiradas dos corpora Brown e WSJ e testando-se em 100 frases aleatórias do corpora Brown.

| | Nome Próprio | | | Átomo Anterior | | | Átomo Posterior | | |
|---|---|---|---|---|---|---|---|---|---|
| | Prec | Cob | Med-F | Prec | Cob | Med-F | Prec | Cob | Med-F |
| WIKI | 0,3784 | 0,4219 | 0,3990 | 0,3454 | 0,1833 | 0,2395 | 0,2950 | 0,1858 | 0,2280 |
| WIKI_P | 0,3987 | 0,4486 | 0,4222 | X | X | X | X | X | X |
| WORDNET | 0,5244 | 0,3090 | 0,3889 | 0,7053 | 0,1365 | 0,2288 | 0,6192 | 0,1035 | 0,1774 |
| MAIUSCULAS | 0,5973 | 0,7313 | 0,6575 | 0,3300 | 0,1812 | 0,2339 | 0,2975 | 0,1650 | 0,2123 |
| NUMEROS | 0,8261 | 0,2241 | 0,3525 | 0,9908 | 0,0108 | 0,0214 | 0,8932 | 0,0142 | 0,0279 |
| ADJS | 0,5375 | 0,1748 | 0,2638 | 0,5288 | 0,1607 | 0,2465 | 0,6433 | 0,1294 | 0,2155 |
| ADVS | 0,6208 | 0,0248 | 0,0478 | 0,7193 | 0,1573 | 0,2581 | 0,8702 | 0,0945 | 0,1705 |
| CONJS | 1,0000 | 0,0000 | 0,0000 | 0,7690 | 0,1165 | 0,2023 | 0,8550 | 0,0892 | 0,1615 |
| DETS | 1,0000 | 0,0000 | 0,0000 | 0,5681 | 0,2053 | 0,3016 | 0,9025 | 0,0225 | 0,0439 |
| NOMES | 0,5856 | 0,6543 | 0,6180 | 0,4179 | 0,2123 | 0,2815 | 0,2792 | 0,2537 | 0,2658 |
| PREPS | 1,0000 | 0,0000 | 0,0000 | 0,5550 | 0,1940 | 0,2875 | 0,6708 | 0,2018 | 0,3102 |
| PRONS | 0,1818 | 0,0868 | 0,1175 | 0,8958 | 0,0847 | 0,1547 | 0,8235 | 0,0413 | 0,0787 |
| VERBOS | 0,6717 | 0,1647 | 0,2645 | 0,7274 | 0,0950 | 0,1681 | 0,5521 | 0,2778 | 0,3696 |
| DESCS | 0,6653 | 0,1962 | 0,3030 | 0,6333 | 0,1495 | 0,2419 | 0,5119 | 0,3358 | 0,4056 |
| ATOMOS | 0,5766 | 0,6801 | 0,6241 | X | X | X | X | X | X |

No que respeita aos nomes próprios as fontes que apresentaram maior impacto foram: a entrada dos nomes próprios nas enciclopédia (no caso a Wikipédia e o WordNet), a presença de maiúsculas e números nos nomes próprios, o comprimento dos nomes próprios e a existência de átomos nos nomes próprios que pertence as classes gramáticais nome comum, adjectivo, verbo ou palavra desconhecida.

No átomo anterior ao nome próprio as fontes que tiveram mais impacto foram: o átomo pertencer a uma das classes gramaticais determinante, preposição, nome comum, advérbio ou adjectivo. Normalmente, palavras destas classes gramaticais antecentem os nomes próprios.

No átomo posterior ao nome próprio as fontes que tiveram mais impacto foram: o átomo pertencer a uma das classes gramaticais verbo, preposição, nome comum, advérbio ou adjectivo. Palavras destas classes gramaticais que precedem os nomes próprios com bastante frequência.

Em relação à classe gramatical palavra desconhecida não podemos especular muito sobre a mesma, porque existe uma frequência baixa de palavras desconhecida no corpora analisado.

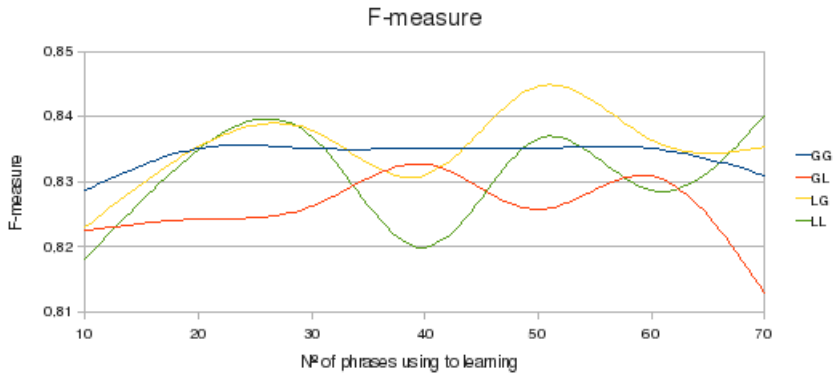
Para optimizar os parâmetros e marcar os nomes próprios foi usada a função heurística:

$$H(I) = P1 * NP\_I(I) + \sum_{i=0}^{i} (P2 * (WIKI(i) + WIKI\_P(i) + WORDNET(i) + NUMEROS(i)) * ATOMOS(i) + P3 * MAIUSCULAS(i) + P4 * (ADVS(i) + CONJS(i) + DETS(i) + PREPS(i) + PRONS(i)))$$

Esta função foi encontrada realizando um conjunto de testes no qual se verificou a importância da combinação das fontes de conhecimento.

Na tabela e no gráficos seguinte podemos ver os valores das métricas obtidos com os marcadores global e local usando os marcadores global e local.

| | MARC | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | GLOBAL | | | | LOCAL | | | | | |
| FRASES | Prec | Cob | Med-F | Corr | Prec | Cob | Med-F | Corr | | |
| 10 | 0,7850 | 0,8773 | 0,8286 | 52 | 0,7805 | 0,8692 | 0,8225 | 50 | GLOBAL | OPT |
| 20 | 0,7889 | 0,8872 | 0,8351 | 54 | 0,7803 | 0,8735 | 0,8243 | 51 | | |
| 30 | 0,7889 | 0,8872 | 0,8351 | 54 | 0,7821 | 0,8757 | 0,8262 | 52 | | |
| 40 | 0,7889 | 0,8872 | 0,8351 | 54 | 0,7906 | 0,8797 | 0,8327 | 53 | | |
| 50 | 0,7889 | 0,8872 | 0,8351 | 54 | 0,7796 | 0,8777 | 0,8258 | 51 | | |
| 60 | 0,7889 | 0,8872 | 0,8351 | 54 | 0,7872 | 0,8798 | 0,8310 | 53 | | |
| 70 | 0,7855 | 0,8822 | 0,8311 | 53 | 0,7715 | 0,8591 | 0,8129 | 50 | | |
| 10 | 0,7797 | 0,8713 | 0,8230 | 47 | 0,7757 | 0,8653 | 0,8181 | 47 | LOCAL | |
| 20 | 0,7905 | 0,8855 | 0,8353 | 54 | 0,7890 | 0,8865 | 0,8349 | 53 | | |
| 30 | 0,7932 | 0,8880 | 0,8379 | 53 | 0,7915 | 0,8880 | 0,8291 | 54 | | |
| 40 | 0,7855 | 0,8822 | 0,8311 | 53 | 0,7742 | 0,8714 | 0,8199 | 50 | | |
| 50 | 0,7994 | 0,8955 | 0,8447 | 53 | 0,7927 | 0,8857 | 0,8366 | 50 | | |
| 60 | 0,7925 | 0,8855 | 0,8364 | 55 | 0,7848 | 0,8778 | 0,8287 | 49 | | |
| 70 | 0,7905 | 0,8855 | 0,8353 | 54 | 0,7950 | 0,8909 | 0,8402 | 51 | | |




Os resultado da tabela e do gráfico mostram que as diferenças na métrica de medida-F entre os marcadores local e global usando os optimizadores local e global não são significativas.

## 7 Conclusões e Trabalho Futuro

Na marcação dos nomes próprios explorar a informação do nome próprio é tão importante com explorar a informação dos átomos que se encontram junto do nome próprio.

As diferenças na métrica de medida-F entre os marcadores local e global usando os optimizadores local e global não foram significativas, existindo no máximo 0,03 valores de diferença. Além disso, usando o marcador local e optimizador local o espaço de pesquisa será reduzido tornando possível em tempo real extrair conclusões que seriam impossíveis de obter apenas com o marcador global e o opmitizador global.

Como perspectiva futura no desenvolvimento do sistema pensamos explorar a marcação de entidades mencionadas.

## References

[1] João Laranjinho and Irene Rodrigues. O impacto de diferentes fontes de conhecimento na marcação de nomes próprios em portugês. In *INFORUM - Simpósio de Informática*, 2010.