

Reconhecimento de Entidades Mencionadas com o XIP: uma colaboração entre o L2F-INESC e a Xerox

C. Hagège

Xerox Research Centre Europe (France)

N. Mamede e J. Baptista

L2f INESC-ID Lisboa (Portugal)

Plano da apresentação

- Contexto geral e metodologia adoptada
- Léxico e pré-processamento
- Gramáticas locais para o REM
- Fases finais do processamento
- Resultados e conclusão

Contexto geral - Metodologia

Contexto Geral e metodologia

REM integrada num contexto mais abrangente de processamento morfossintáctico do Português

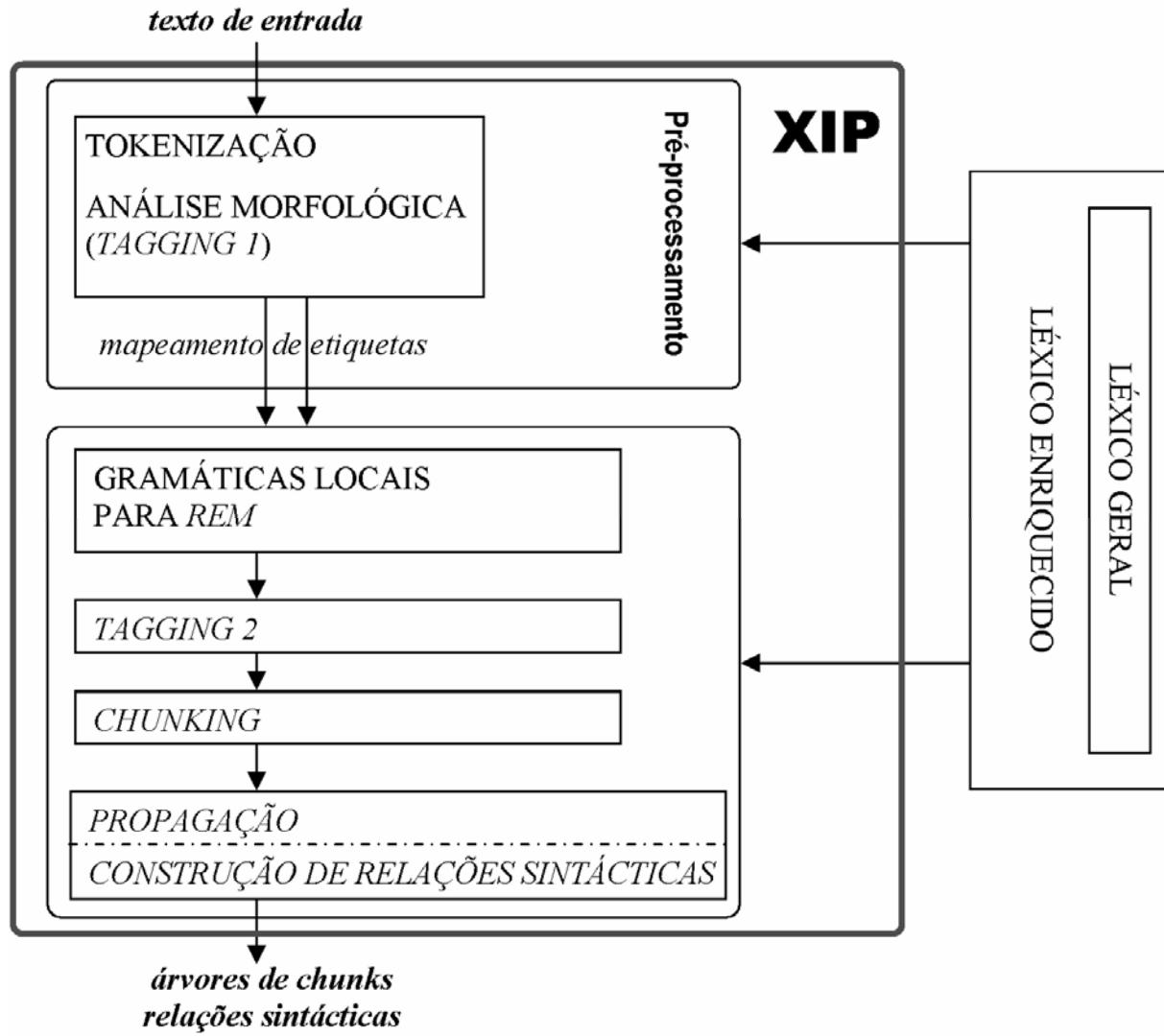
- Processamento geral beneficia do correcto tratamento das EM
- Melhor reconhecimento das EM se estiver disponível conhecimento do contexto (eg. metonímia)

quadro geral : Xerox Incremental Parser (XIP)

[Ait et Al. 02] Salah Ait-Mokhtar, Jean-Pierre Chanod, Claude Roux (2002)
“Robustness beyond shallowness: incremental dependency parsing”. Special issue of the NLE Journal (2002).

- Processamento incremental
- Várias línguas processadas em estado de desenvolvimento variado
- Formalismo rico, que permite expressar conhecimento linguístico que vai da desambiguação ao cálculo de relações sintácticas entre constituintes passando pela construção de domínios sintácticos.
- Linguagem de *scripting* integrada para tarefas extra-lingüísticas
- API Python e API Java

para o Português



Léxico e pré-processamento

Léxico

Dois tipos de léxicos:

- Léxico oriundo do pré-processamento que pode ser ou não enriquecido por novos traços no XIP

arcebispo: noun += [cargo:+] .

- Novas entradas lexicais

Herbert += [people:+, individual:+, firstname:+] .

IMPORTANTE:

A tokenização do pré-processamento não pode ser mudada no léxico do XIP

Pré-processamento

- Consiste na tokenização, análise morfológica e primeira fase de desambiguação.
- Resultado integrado no XIP
- Fase de *mapping* necessária
(à partida qualquer ferramenta de pré-processamento pode ser integrada no XIP)

Dois pré-processamentos diferentes
mas **um só analisador** para EM e sintaxe.

Disambiguação

Exemplo 1

Regra de TAGGING específica (antes da aplicação das regras locais)

```
verb<lemma:podar>, verb<lemma:poder> =  
          verb<lemma:poder> | verb[inf:+] | .
```

Exemplo 2

Regra de TAGGING específica desenvolvida para REM

```
noun[maj:+,surface:Natal] %=  
          | noun[denot_time:+], prep[lemma:de], art |  
noun[one_day=+,maj=+,proper=+] .
```

Gramáticas locais para o REM

Gramáticas locais

Regras de reescritas contextuais

- Juntar vários *tokens* para criar um novo nó
exemplos: *Sargento-mor*, *Cônsul Honorário*

```
1> noun [cargo=+, mwe=+, people=+] @=
? [cargo, maj], (punct [hifen]), adj [lemma:"honorário",
maj]; adj [lemma:mor, maj] .
```

- Uso do contexto local para tipificar e delimitar EM.
exemplo: *governo de Lisboa*

```
1>NOUN [org=+, institution=+] @=
| ? [lemma:governo, maj:~], prep [lemma:de], (art) |
? [location] .
```

Fases finais do processamento

Chunking e Dependências

XIP UI - 0.2.26 - XIP: Xerox Incremental Parser 9.63 "build 55" (2000-2007) -

Project Help

NE_PORT

File Edit Grammar Corpus Run Tools Input

Corpus Grammar

Rule Map

Files

- dependency1.xip
- LGTime.xip
- chunker.xip
- dependencyLast.xip
- dependency8.xip
- LGNumber_xerox.xip
- LGPeople.xip
- dependency4.xip
- LGCulture.xip
- LG0.xip
- dependency0.xip
- LGOrg.xip
- dependency4entit.xip
- LGLast.xip
- LGEVENT.xip
- dependency7.xip
- LGElectronic.xip
- TermLocation.xip
- disamb4entit.xip
- LGMeasure.xip
- dependency3.xip
- LGLocation.xip

Filters : 99

Results

input Total: 1 file(s) error(s): 0 Size: 1 Kb

Size: 1 Kb Encoding: UTF-8 Parent: Default_corpus File: /tmp/input_64715.txt

TOP

```
+-----+-----+-----+
|       |       |       |
|   NP  |   VF  |   PP  |
|   +---+   +---+   +---+
|   ART |   VERB |   PREP |
|   +---+   +---+   +---+
|   A    |   NOUN |   NOUN |
|   +---+   +---+   +---+
|   NOUN |   NOUN |   vivia |
|   +---+   +---+   +---+
|   Joaninha |   Sampaio |   na |
|               |               |   o |
|               |               |   Lourinhã |
|               |               |   . |
```

O> A Joaninha Sampaio vivia na o Lourinhã.
NE_INDIVIDUAL_PEOPLE_[1524] (Joana Sampaio)
NE_LOCAL_CITY_ADMIN_AREA_[1530] (Lourinhã)
PREPD_[1769] (Lourinhã,em)
DETD_[1780] (Joana Sampaio,o)
DETD_[1780] (Lourinhã,o)
VDOMAIN_[1819] (viver,viver)
MOD_POST_[2491] (viver,Lourinhã)
SUBJ_PRE_[2511] (viver,Joana Sampaio)
MAIN_[2641] (viver)
O>TOP {NP{o NOUN{Joana Sampaio}}} VF{viver} PP{em o Lourinhã} .}

Parse (5)

Output Input

13

A Joaninha Sampaio vivia na Lourinhã.

0:37 | UTF-8 : input.txt

Refinamento com uso da sintaxe

Exemplo

Portugal, que votou esta lei ...

```
if ( ^NE[local:+,admin_area:+] (#1) &
  ( SUBJ(?[lemma:promulgar],#1) |
    SUBJ(?[lemma:votar],#1)
  )
)
NE[features=~,org=+,administration=+] (#1)
```

Propagação (1)

Exemplo

Um capitão norueguês chamado **Trygve Petersen** conduziu o Mira de novo a Portugal . . .

<frase intermédia>

. . . **Petersen** não trazia carga nenhuma.

NE_INDIVIDUAL_PEOPLE(**Trygve Petersen**)'|

Propagação (2)

1º) MARCAR

Um capitão norueguês chamado **Trygve Petersen** conduziu o Mira de novo a Portugal

```
noun#1[people,individual]{?*,noun#2[title:~,location:  
:~,org:~,initial:~,maj:+],  
?*,  
noun#3[last,title:~,location:~,initial:~,maj:+]} |  
if (NE[people](#1))  
{ PERSON##2=1; PERSON##3=1; }
```

Propagação (3)

2) RESTITUIR

[Trygve Petersen] . . . **Petersen** não trazia carga nenhuma.

```
| noun#1[toutmaj:~,maj:+] |
  if ( PERSON##1:1 & ~NE[people] (#1) )
    NE[people=+,individual=+,propag=+] (#1)
```



NE_INDIVIDUAL_PEOPLE_PROPAG (**Petersen**)

Resultados e conclusão

Resultados e Conclusão

- Resultados muito encorajadores
 - para uma primeira participação
 - e com um sistema recente
- Não considerámos todos os tipos do HAREM (COISA, ABSTRACCAO)
- Ponto forte no TEMPO (clássico e específico)
- Foi privilegiada a *precisão* e não a *abrangência*
- Muito ficou por fazer
 - (integração de léxicos, propagação, etc.)
- mas há espaço para melhoramento . . .

Obrigado !

C. Hagège

Xerox Research Centre Europe (France)

N. Mamede e J. Baptista

L2f/INESC-ID Lisboa (Portugal)

