

Extracção e Generalização de Recursos de Tradução com base em Dicionários Probabilísticos

Alberto Manuel Brandão Simões
ambs@di.uminho.pt

Orientação: *José João Almeida*
jj@di.uminho.pt

SDDI'07
21 a 23 de Fevereiro

- 1 Contextualização e Motivação
- 2 Dicionários Probabilísticos de Tradução
- 3 Extracção de Terminologia Bilingue
- 4 Generalização de Terminologia Bilingue
- 5 Paralelização

- 1 Contextualização e Motivação
- 2 Dicionários Probabilísticos de Tradução
- 3 Extracção de Terminologia Bilingue
- 4 Generalização de Terminologia Bilingue
- 5 Paralelização

- 1 Contextualização e Motivação
- 2 Dicionários Probabilísticos de Tradução
- 3 Extracção de Terminologia Bilingue
- 4 Generalização de Terminologia Bilingue
- 5 Paralelização

- 1 Contextualização e Motivação
- 2 Dicionários Probabilísticos de Tradução
- 3 Extracção de Terminologia Bilingue
- 4 Generalização de Terminologia Bilingue
- 5 Paralelização

- 1 Contextualização e Motivação
- 2 Dicionários Probabilísticos de Tradução
- 3 Extracção de Terminologia Bilingue
- 4 Generalização de Terminologia Bilingue
- 5 Paralelização

Parte I

Contextualização e Motivação

Tradução Assistida por Computador

- abstrai os formatos dos documentos a traduzir;
- ajuda a **reutilizar traduções**;
- várias ferramentas disponíveis;
SDL Trados, Star Transit, Déjà vu, Omega-T, TRANSBey

Processo de Tradução

- 1 o programa **divide** o texto original em segmentos;
- 2 a tradução é realizada para cada segmento:
 - 1 verificar se o segmento está na **memória de tradução** (ou algum segmento razoavelmente semelhante)
 - 2 o tradutor determina se vai **usar, editar ou ignorar** essa tradução;
 - 3 o programa **guarda o par de segmentos**;

Tradução Assistida por Computador

- abstrai os formatos dos documentos a traduzir;
- ajuda a **reutilizar traduções**;
- várias ferramentas disponíveis:
SDL Trados, Star Transit, Déjà vu, Omega-T, TRANSBey

Processo de Tradução

- 1 o programa **divide** o texto original em segmentos;
- 2 a tradução é realizada para cada segmento:
 - 1 verificar se o segmento está na **memória de tradução** (ou algum segmento razoavelmente semelhante)
 - 2 o tradutor determina se vai **usar, editar ou ignorar** essa tradução;
 - 3 o programa **guarda o par de segmentos**;

Tradução Baseada em Regras

- usam dicionários bilingues;
- usam regras de pré-edição;
- usam regras de transferência;
- usam regras de pós-edição;

Tradução Baseada em Exemplos

- usam dicionários bilingues;
- usam traduções já realizadas:
 - divididas em segmentos;
 - generalizadas;

e outras...

Tradução Baseada em Dados; Tradução Baseada em Frases; ...



Tradução Baseada em Regras

- usam dicionários bilingues;
- usam regras de pré-edição;
- usam regras de transferência;
- usam regras de pós-edição;

Tradução Baseada em Exemplos

- usam dicionários bilingues;
- usam traduções já realizadas:
 - divididas em segmentos;
 - generalizadas;

e outras...

Tradução Baseada em Dados; Tradução Baseada em Frases; ...

Tradução Baseada em Regras

- usam dicionários bilingues;
- usam regras de pré-edição;
- usam regras de transferência;
- usam regras de pós-edição;

Tradução Baseada em Exemplos

- usam dicionários bilingues;
- usam traduções já realizadas:
 - divididas em segmentos;
 - generalizadas;

e outras. . .

Tradução Baseada em Dados; Tradução Baseada em Frases; . . .



Estes resultados constituem a base do Programa Europeu de defesa do Mar de Barents e, por esse motivo, peço-lhe que analise um projecto de carta que lhe expõe os factos mais importantes, e que, de acordo com as decisões do Parlamento, torne clara esta posição na Rússia.

No entanto, somos também da opinião de que deveria haver um debate sobre esta estratégia da comissão que seguisse um procedimento ordenado, e não só com base numa declaração oral pronunciada aqui no Parlamento Europeu, mas também com base num documento que seja decidido na comissão e que apresente uma descrição deste programa para um período de cinco anos.

These findings form the basis of the European Programmes to protect the Barents Sea, and that is why I would ask you to examine a draft letter setting out the most important facts and to make Parliament's position, as expressed in the resolutions which it has adopted, clear as far as Russia is concerned.

We believe, however, that the commission's strategic plan needs to be debated within a proper procedural framework, not only on the basis of an oral statement here in the European Parliament, but also on the basis of a document which is adopted in the commission and which describes this programme over the five-year period .

Estes resultados constituem a base do Programa Europeu de defesa do Mar de Barents e, por esse motivo, peço-lhe que analise um projecto de carta que lhe expõe os factos mais importantes, e que, de acordo com as decisões do Parlamento, torne clara esta posição na Rússia.

No entanto, somos também da opinião de que deveria haver um debate sobre esta estratégia da comissão que seguisse um procedimento ordenado, e não só com base numa declaração oral pronunciada aqui no Parlamento Europeu, mas também com base num documento que seja decidido na comissão e que apresente uma descrição deste programa para um período de cinco anos.

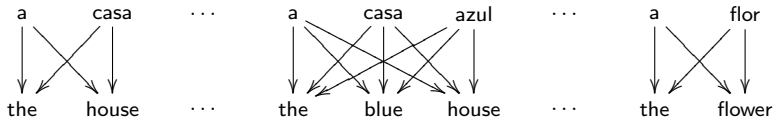
These findings form the basis of the European Programmes to protect the Barents Sea, and that is why I would ask you to examine a draft letter setting out the most important facts and to make Parliament's position, as expressed in the resolutions which it has adopted, clear as far as Russia is concerned.

We believe, however, that the commission's strategic plan needs to be debated within a proper procedural framework, not only on the basis of an oral statement here in the European Parliament, but also on the basis of a document which is adopted in the commission and which describes this programme over the five-year period .

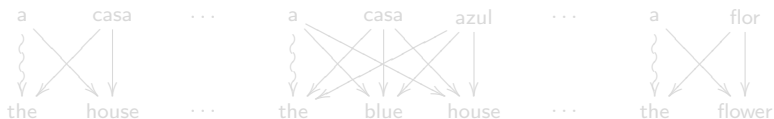
Parte II

Dicionários Probabilísticos de Tradução

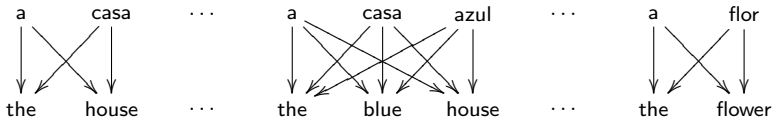
Co-ocorrências iniciais:



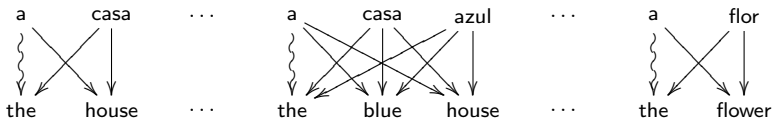
Usar as correlações em maior número:



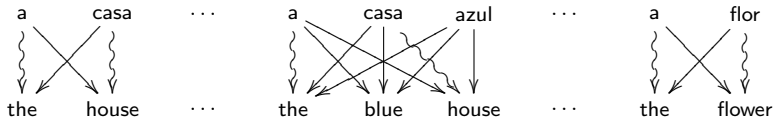
Co-ocorrências iniciais:



Usar as correlações em maior número:



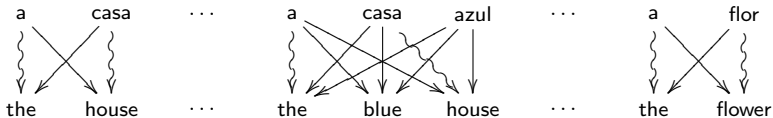
Continuar a inferir resultados:



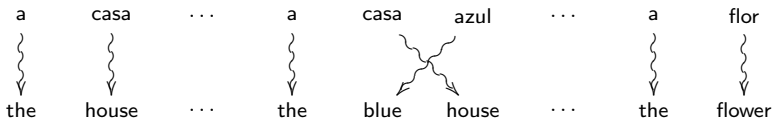
Finalmente:



Continuar a inferir resultados:



Finalmente:



Parte III

Extracção de Terminologia Bilingue



Motivação

- a ordem das palavras no alinhamento pode variar bastante;
- no entanto, podemos prever algumas mudanças (padrões):
 - dependentes de pares de língua;
 - permitem extrair entradas terminológicas;
 - permitem extrair propriedades de exemplos;

Metodologia

- Definir uma DSL:
 - para descrever os padrões;
 - para descrever restrições sobre os padrões;
 - para descrever as propriedades que se inferem;
 - simples para que de forma rápida se crie padrões para qualquer par de língua.

Motivação

- a ordem das palavras no alinhamento pode variar bastante;
- no entanto, podemos prever algumas mudanças (padrões):
 - dependentes de pares de língua;
 - permitem extrair entradas terminológicas;
 - permitem extrair propriedades de exemplos;

Metodologia

- Definir uma DSL:
 - para descrever os padrões;
 - para descrever restrições sobre os padrões;
 - para descrever as propriedades que se inferem;
 - simples para que de forma rápida se crie padrões para qualquer par de língua.

	Jogos	Oímpicos
Olimpic		X
Games	X	

Descrita como:

$$[ABBA] \ A \ B = B \ A$$

Formalmente...

$$T(A \cdot B) = T(B) \cdot T(A)$$

	índice	de	desenvolvimento	humano
human				X
development			X	
index	X			

Descrita como:

[IDH] I "de" D H = H D I

Formalmente...

$$T(I \cdot \text{"de"} \cdot D \cdot H) = T(H) \cdot T(D) \cdot T(I)$$

	protocolo	de	transferência	de	ficheiros
file					X
transfer			X		
protocol	X				

Descrita como:

[FTP] P "de" T "de" F = F T P

Formalmente...

$$T(P \cdot \text{"de"} \cdot T \cdot \text{"de"} \cdot F) = T(F) \cdot T(T) \cdot T(P)$$

	ponto	de	vista	neutro
neutral				X
point	X			
of		Δ		
view			X	

Descrita como:

[POV] P "de" V N = N P "of" V

Formalmente...

$$\mathcal{T}(P \cdot \text{"de"} \cdot V \cdot N) = \mathcal{T}(N) \cdot \mathcal{T}(P) \cdot \text{"of"} \cdot \mathcal{T}(V)$$



- a maior parte destes padrões definem nominais;
- no entanto também podem fazer **matching** com outras sequências de palavras;
- mas com um pouco mais de informação nos padrões será possível limitar a sua aplicabilidade;
- **definição de padrões com restrições:**
 - restrições morfológicas;
 - predicados genéricos (Perl);
- **definição de regras de inferência;**

Restrições Morfológicas

Definição de propriedades morfológicas que devem **ser válidas** para a expressão em causa.

$$[ABBA] A B[CAT<-adj] = B[CAT<-adj] A$$

Predicados Genéricos

Definição de predicados genéricos, definindo **funções Perl** sobre palavras ou conjunto de palavras.

$$[ABBA] A B.is_adj = B.is_adj A$$

Inferência

Realizar **inferência de propriedades** (normalmente morfológicas) a partir das expressões encontradas.

$$[ABBA] A B[CAT->adj] = B[CAT->adj] A$$

Restrições Morfológicas

Definição de propriedades morfológicas que devem **ser válidas** para a expressão em causa.

$$[ABBA] A B[CAT<-adj] = B[CAT<-adj] A$$

Predicados Genéricos

Definição de predicados genéricos, definindo **funções Perl** sobre palavras ou conjunto de palavras.

$$[ABBA] A B.is_adj = B.is_adj A$$

Inferência

Realizar **inferência de propriedades** (normalmente morfológicas) a partir das expressões encontradas.

$$[ABBA] A B[CAT->adj] = B[CAT->adj] A$$

Restrições Morfológicas

Definição de propriedades morfológicas que devem **ser válidas** para a expressão em causa.

$$[ABBA] A B[CAT<-adj] = B[CAT<-adj] A$$

Predicados Genéricos

Definição de predicados genéricos, definindo **funções Perl** sobre palavras ou conjunto de palavras.

$$[ABBA] A B.is_adj = B.is_adj A$$

Inferência

Realizar **inferência de propriedades** (normalmente morfológicas) a partir das expressões encontradas.

$$[ABBA] A B[CAT->adj] = B[CAT->adj] A$$

	discussion	about	alternative	sources	of	financing	for	the	europaean	radical	alliance	.
discussão	44	0	0	0	0	0	0	0	0	0	0	0
sobre	0	11	0	0	0	0	0	0	0	0	0	0
fontes	0	0	0	74	0	0	0	0	0	0	0	0
de	0	3	0	0	27	0	6	3	0	0	0	0
financiamento	0	0	0	0	0	56	0	0	0	0	0	0
alternativas	0	0	23	0	0	0	0	0	0	0	0	0
para	0	0	0	0	0	0	28	0	0	0	0	0
a	0	1	0	0	1	0	4	33	0	0	0	0
aliança	0	0	0	0	0	0	0	0	0	0	65	0
radical	0	0	0	0	0	0	0	0	0	80	0	0
européia	0	0	0	0	0	0	0	0	59	0	0	0
.	0	0	0	0	0	0	0	0	0	0	0	80

39214 = comunidades europeias !=ABBA!= european communities
32850 = jornal oficial !=ABBA!= official journal
32832 = parlamento europeu !=ABBA!= european parliament
32730 = união europeia !=ABBA!= european union
31650 = comunidade europeia !=ABBA!= european community
15602 = países terceiros !=ABBA!= third countries
[...]
3614 = livro verde !=ABBA!= green paper
3520 = saúde pública !=ABBA!= public health
3434 = direito comunitário !=ABBA!= community law
3243 = conselho europeu !=ABBA!= european council
3227 = nível comunitário !=ABBA!= community level
3179 = comité permanente !=ABBA!= standing committee
3038 = nomenclatura combinada !=ABBA!= combined nomenclature
[...]
1 = órgãos orçamentais !=ABBA!= budgetary organs
1 = órgãos relevantes !=ABBA!= relevant bodies
1 = óvulos de equino !=A!= equine ova
1 = óxido de albendazole !=A!= albendazole oxide
1 = óxido de cádmio !=A!= cadmium oxide
1 = óxido de estireno !=A!= styrene oxide

21007 união europeia => european union
9301 parlamento europeu => european parliament
4171 direitos humanos => human rights
3504 estados unidos => united states
2353 mercado interno => internal market
1911 posição comum => common position
1826 países candidatos => candidate countries
1776 comissão europeia => european commission
1708 conselho europeu => european council
1629 saúde pública => public health
1558 direitos fundamentais => fundamental rights
1546 nações unidas => united nations
1337 países terceiros => third countries
1294 conferência intergovernamental => intergovernmental conference
1258 fundos estruturais => structural funds

729 plano de acção => action plan
722 conselho de segurança => security council
680 processo de paz => peace process
582 mercado de trabalho => labour market
580 pena de morte => death penalty
492 pacto de estabilidade => stability pact
431 política de defesa => defence policy
353 acordo de associação => association agreement
348 protocolo de quioto => kyoto protocol
343 programa de acção => action programme
259 branqueamento de capitais => money laundering
258 comité de conciliação => conciliation committee
241 política de concorrência => competition policy
226 processo de conciliação => conciliation procedure
217 requerentes de asilo => asylum seekers

- 93 tribunal de justiça europeu => european court of justice
- 51 tribunal de contas europeu => european court of auditors
- 33 fontes de energia renováveis => renewable sources of energy
- 27 ponto de vista ambiental => environmental point of view
- 26 ponto de vista económico => economic point of view
- 21 ponto de vista jurídico => legal point of view
- 20 declaração de fiabilidade positiva => positive statement of assurance
- 18 ponto de vista político => political point of view
- 13 ponto de vista técnico => technical point of view
- 10 ponto de vista institucional => institutional point of view
- 9 ponto de vista orçamental => budgetary point of view
- 8 sistema de preferências generalizadas => generalised system of preferences
- 8 método de coordenação aberto => open method of coordination
- 7 ponto de vista social => social point of view
- 7 ponto de vista democrático => democratic point of view



- EuroParl PT-EN: 1 000 000 TUs
- 700 000 unidades de tradução processadas
- 578 103 ocorrências de padrões
- 139 781 exemplos diferentes
- 103 617 exemplos depois de filtrados (**stop words, ruído,...**)
- 77 497 ex. com a regra $A B = B A$ (938/2/1) **(86%)**
- 12 694 ex. com a regra $A \text{ "de" } B = B A$ (204/2/1) **(95%)**
- 7 700 ex. com a regra $A B C = C B A$ (40/1/1) **(93%)**
- 3 336 ex. com a regra $H \text{ "de" } D H = H D I$ (21/1/1) **(100%)**
- 1 466 ex. com a regra $A B C = C A B$ (4/1/1) **(40%)**
- 564 ex. com a regra $P \text{ "de" } V N = N P \text{ "of" } V$ (6/1/1) **(98%)**
- 360 ex. com a regra $P \text{ "de" } T \text{ "de" } F = F T P$ (3/1/1) **(96%)**

Parte IV

Generalização de Terminologia Bilingue

Motivação

- muitas entradas terminológicas \Rightarrow muito espaço em disco;
- entradas terminológicas estão completamente instanciadas;
 - \Rightarrow reusabilidade limitada;

Abordagem

- calcular classes de entidades:
 - classes de entradas numéricas;
 - classes de entidades mencionadas;
 - classes de palavras;

Motivação

- muitas entradas terminológicas \Rightarrow muito espaço em disco;
- entradas terminológicas estão completamente instanciadas;
 - \Rightarrow reusabilidade limitada;

Abordagem

- calcular classes de entidades:
 - classes de entradas numéricas;
 - classes de entidades mencionadas;
 - classes de palavras;

399	às <code>_horaA_</code>	<code>_horaB_</code>
187	orçamento de <code>_ano_</code>	<code>_ano_ budget</code>
136	<code>_int_ euros</code>	<code>eur _int_</code>
135	<code>_int_ euros</code>	<code>eur _int_</code>
127	directiva de <code>_ano_</code>	<code>_ano_ directive</code>
51	orçamento <code>_ano_</code>	<code>_ano_ budget</code>
46	<code>_int_ de setembro</code>	<code>september _int_</code>
31	partir de <code>_ano_</code>	<code>_ano_ onwards</code>
29	convenção de <code>_ano_</code>	<code>_ano_ convention</code>
26	eleições de <code>_ano_</code>	<code>_ano_ elections</code>
25	período <code>_ano_-ano_</code>	<code>_ano_-ano_ period</code>
25	<code>_int_ dólares</code>	<code>usd _int_</code>
24	relatório de <code>_ano_</code>	<code>_ano_ report</code>
21	convenção de genebra de <code>_ano_</code>	<code>_ano_ geneva convention</code>
17	período de <code>_ano_-ano_</code>	<code>_ano_-ano_ period</code>



2	povo	português	portuguese	people
2	povo	paraguaio	paraguyan	people
2	povo	nigeriano	nigerian	people
2	povo	mexicano	mexican	people
2	povo	marroquino	moroccan	people
2	povo	mapuche	mapuche	people
2	povo	indígena	indigenous	people
2	povo	holandês	dutch	people
2	povo	húngaro	hungarian	people
2	povo	hmong	hmong	people
2	povo	guatemalteco	guatemalan	people

generalizando...

povo	X:	gentílico(X)	T(X)	people
governo	X:	gentílico(X)	T(X)	governo

- Para cada padrão:
 - fixar um (ou mais elementos);
 - por exemplo, em $AB=BA$ fixar A (substantivo);
 - calcular conjunto de Bs possíveis;
 - conversão em multi-sets;
 - estatísticas sobre multi-sets;

```
'ácido' => [ 'clorídrico (hydrochloric acid)',  
             'sulfúrico (sulphuric acid)',  
             'acético (acetic acid)',  
             'fólico (folic acid)',  
             'cítrico (citric acid)',  
             'nítrico (nitric acid)',  
             'tartárico (tartaric acid)',  
             'benzóico (benzoic acid)',  
             'fórmico (formic acid)',  
             'málico (malic acid)',  
             'sulfúrico (sulfuric acid)',  
             'erúcico (erucic acid)',
```

...

```
'livro' => [ 'verde (green paper)',  
            'branco (white paper)',  
            'azul (blue paper)',  
            'aberto (open book)',  
            'azul (blue book)',  
            'branco (white book)',  
            'laranja (orange book)',  
            'vermelho (red book)'
```

...


```
ácido X.acidClass = X.acidClass acid
```

em que X.acidClass contém:

clorídrico = hydrochloric

sulfúrico = sulphuric

acético = acetic

fólico = folic

cítrico = citric

nítrico = nitric

tartárico = tartaric

benzóico = benzoic

fórmico = formic

málico = malic

sulfúrico = sulfuric

erúcico = erucic

...

Parte V

Paralelização

- processo de alinhamento e de extracção de terminologia demora **demasiado tempo** (uma semana...);
- afinação de parâmetros de algoritmos obriga ao **re-cálculo** de todo o processo;
- esperar uma semana é impensável;
- **paralelizar o processo** ao nível da instrução está fora do âmbito do doutoramento...
- ...no entanto a paralelização **ao nível da aplicação** não é tão dispendiosa;
- mas é necessário um método expedito de **especificar formalmente dependências** entre as várias aplicações que constituem o processo de alinhamento e extracção de terminologia;

- utilizar uma linguagem conhecida para especificar dependências: **makefile**;
- dependências entre aplicações e não dependência de ficheiros;
- açúcar sintáctico extra para alguns detalhes específicos;
- acções semânticas em Bash (como as makefiles)
- ou em Perl (linguagem de eleição).
- regras paramétricas:
 - instanciar regras em tempo de execução;
 - facilitar paralelismo da mesma aplicação;
 - facilitar execução com parâmetros diferentes;
 - detalhes no próximo dia 28...

```

codify: (20:00:00)
    nat-codify -id=EurLex -tmx eurlex.tmx
    i <- sub{ $nr = 'cat EurLex/nat.cnf |grep nr-chunks|cut -f 2 -d "="; printf("%03d\n",$_) for (1..$nr)
    j <- sub{ print("$_\n") for (2..4); }

ngramsA$j: codify (20:00:00)
    for a in @i; do nat-ngrams -n $j EurLex/source.${a}.crp /tmp/source.$j.ngrams; done
    mv /tmp/source.$j.ngrams EurLex/source.$j.ngrams

ngramsB$j: codify (20:00:00)
    for a in @i; do nat-ngrams -n $j EurLex/target.${a}.crp /tmp/target.$j.ngrams; done
    mv /tmp/target.$j.ngrams EurLex/target.$j.ngrams

ngramsA: ngramsA$j (20:00:00)
    nat-joinGrams -tmp=/tmp -s EurLex

ngramsB: ngramsB$j (20:00:00)
    nat-joinGrams -tmp=/tmp -t EurLex

ngrams: ngramsA ngramsB (20:00:00)
    echo 'n-grams=1' >> EurLex/nat.cnf

initmat$i: codify (20:00:00)
    nat-initmat EurLex/source.$i.crp EurLex/target.$i.crp EurLex/mat.$i.in

ipfp$i: initmat$i (20:00:00)
    nat-ipfp 5 EurLex/source.$i.crp EurLex/target.$i.crp EurLex/mat.$i.in EurLex/mat.$i.out
    rm -f EurLex/mat.$i.in

```



Instanciação de uma variável

```
foo: ...
...
i <- sub{ $nr = 'cat Eu...'; printf("%03d\n",$_) for (1..$nr); }
```

Regra paramétrica

```
initmat$i: codify (20:00:00)
  initmat EurLex/source.$i.crp EurLex/target.$i.crp EurLex/mat.$i.in
```

Regra expandida

```
initmat001: codify (20:00:00)
  initmat EurLex/source.001.crp EurLex/target.001.crp EurLex/mat.001.in

initmat002: codify (20:00:00)
  initmat EurLex/source.002.crp EurLex/target.002.crp EurLex/mat.002.in

initmat003: codify (20:00:00)
  initmat EurLex/source.003.crp EurLex/target.003.crp EurLex/mat.003.in

...
```



Instanciação de uma variável

```
foo: ...  
  ...  
  i <- sub{ $nr = 'cat Eu...'; printf("%03d\n",$_) for (1..$nr); }
```

Regra paramétrica

```
initmat$i: codify (20:00:00)  
  initmat EurLex/source.$i.crp EurLex/target.$i.crp EurLex/mat.$i.in
```

Regra expandida

```
initmat001: codify (20:00:00)  
  initmat EurLex/source.001.crp EurLex/target.001.crp EurLex/mat.001.in  
  
initmat002: codify (20:00:00)  
  initmat EurLex/source.002.crp EurLex/target.002.crp EurLex/mat.002.in  
  
initmat003: codify (20:00:00)  
  initmat EurLex/source.003.crp EurLex/target.003.crp EurLex/mat.003.in  
...
```

Instanciação de uma variável

```
foo: ...  
  ...  
  i <- sub{ $nr = 'cat Eu...'; printf("%03d\n",$_) for (1..$nr); }
```

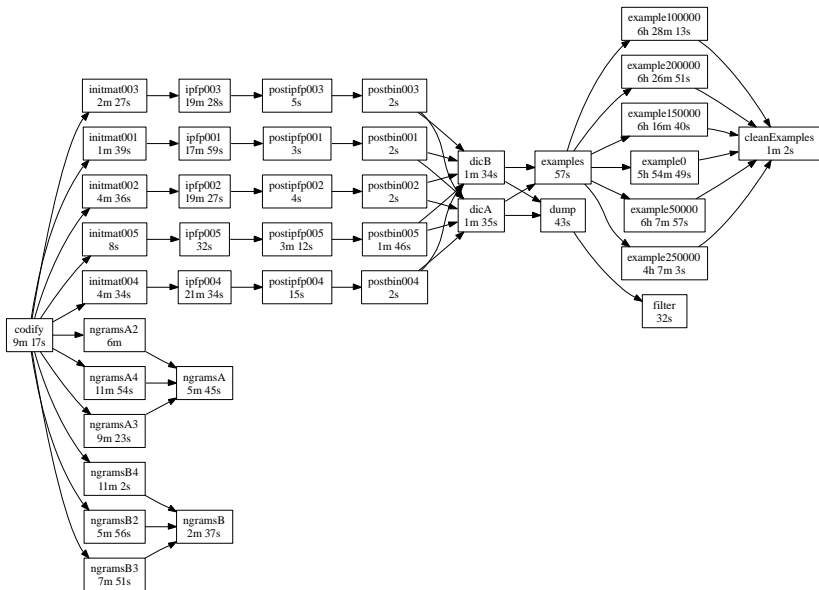
Regra paramétrica

```
initmat$i: codify (20:00:00)  
  initmat EurLex/source.$i.crp EurLex/target.$i.crp EurLex/mat.$i.in
```

Regra expandida

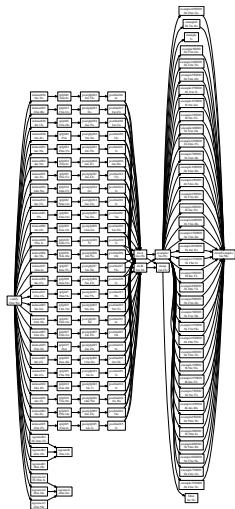
```
initmat001: codify (20:00:00)  
  initmat EurLex/source.001.crp EurLex/target.001.crp EurLex/mat.001.in  
  
initmat002: codify (20:00:00)  
  initmat EurLex/source.002.crp EurLex/target.002.crp EurLex/mat.002.in  
  
initmat003: codify (20:00:00)  
  initmat EurLex/source.003.crp EurLex/target.003.crp EurLex/mat.003.in  
...
```


Paralelização da Extração de Terminologia



Na verdade, os processos típicos que têm sido executados, compreendem:

- mais de 150 trabalhos;
- mais de 25 níveis de paralelismo;
- mais de 16 milhões de exemplos extraídos;
- mais de 1 milhão de entradas terminológicas;
- mais de 10 GB de informação produzida;
- tempo crítico de execução: 4h 30m
- **bottleneck**: acesso a disco



Parte VI

Concluindo...

- processamento de textos paralelos
(Simões & Almeida 2006b)
- extracção de dicionários de tradução
(Simões 2004)
- filtragem de memórias de tradução
(Almeida & Simões 2007)
- extracção de terminologia bilingue
(Simões & Almeida 2007)
- extracção de exemplos de tradução
(Simões & Almeida 2006a)
- generalização de terminologia e exemplos
(Simões & Almeida 2007)
- paralelização
(Simões & Fonseca & Almeida 2007?)

- Simões & Fonseca & Almeida 2007** Alberto Simões, Rúben Fonseca, and José João Almeida. Makefile::Parallel dependency specification language. **Unpublished**, 2007.
- Simões & Almeida 2007** Alberto Simões and José João Almeida. Using alignment patterns for bilingual terminology extraction. In *The 31st Annual Conference of the German Classification Society on Data Analysis, Machine Learning, and Applications*, 2007. **Forthcoming**.
- Almeida & Simões 2007** José João Almeida and Alberto Simões. XML::TMX – processamento de memórias de tradução de grandes dimensões. In José Carlos Ramalho, João Correia Lopes, and Luís Carriço, editors, *XATA 2006 – 5ª Conferência Nacional em XML, Aplicações e Tecnologias Aplicadas*, pages 83–93, February 2007.
- Simões & Almeida 2006b** Alberto Simões and J. João Almeida. NatServer: a client-server architecture for building parallel corpora applications. *Procesamiento del Lenguaje Natural*, 37:91–97, September 2006.
- Simões & Almeida 2006a** Alberto Simões and J. João Almeida. Combinatory examples extraction for machine translation. In Jan Tore Lønning and Stephan Oepen, editors, *11th Annual Conference of the European Association for Machine Translation*, pages 27–32, Oslo, Norway, 19–20, June 2006.
- Simões 2004** Alberto Manuel Brandão Simões. Parallel corpora word alignment and applications. Master's thesis, Escola de Engenharia - Universidade do Minho, 2004.