PAPEL

Palavras Associadas Porto Editora Linguateca

Utilização do (analisador sintáctico) PEN para extracção de informação das definições de um dicionário

Hugo Gonçalo Oliveira, Paulo Gomes
Linguateca, pólo de Coimbra, DEI - FCTUC, CISUC

Novembro 2008

Índice

1 Introdução

Os dois primeiros relatórios do PAPEL relataram o estado da arte relativamente a recursos semelhantes àquele que pretendemos construir [6] e a trabalho feito ao nível da extracção de relações a partir de dicionários electrónicos [7].

Este terceiro relatório começa por mostrar os ngramas mais frequentes no Dicionário da Língua Portugesa da Porto Editora [1], uma fonte importante para determinar que padrões devem ser utilizados na extracção de relações a partir do mesmo. É depois apresentada a ferramenta que serve de base à extracção de informação a partir das relações do(de) dicionário(s), o analisador sintáctico PEN, e explicado de que forma é utilizado para atingir o nosso objectivo.

2 Ngramas mais frequentes no Dicionário

De forma a detectar os padrões mais frequentes no Dicionário, foi construída uma tabela com os ngramas que ocorrem no texto das suas definições, distribuídos por posição na definição, categoria gramatical da definição onde ocorrem e frequência de ocorrência em todas as definições.

Esta grande quantidade de informação serve de auxílio à compreensão da estrutura das definições e permite-nos tirar partido de vários padrões característicos, indicadores de determinadas relações no Dicionário.

Através de uma análise empírica destes dados, complementada com a observação de definições onde os ngramas ocorrem é possível tirar várias conclusões acerca dos padrões textuais que faz sentido ou não utilizar para a extracção de relações.

A título de exemplo, as Tabelas 1, 2, 3 e 4 mostram os ngramas mais frequentes a iniciar definições de palavras de classes gramaticais abertas.

Ngrama	**Frequência**
acto ou efeito de	3851
pessoa que	1386
indivíduo	1247
aquele que	1148
conjunto de	1004
parte	1052
o que	875
espécie de	798
qualidade de	777
qualidade do que é	663

Tabela 1: Ngramas frequentes a iniciar definições de substantivos.

Ngrama	**Frequência**
fazer	1680
tornar	1359
dar	1188
tirar	744
pôr	674
ter	467

Tabela 2: Ngramas frequentes a iniciar definições de verbos.

Ngrama	Frequência
que tem	2698
diz-se	2066
que ou aquele que	1392
relativo a	1236
relativo à	1162
relativo ao	725
relativo ou pertencente	647
que ou que	527
que diz respeito	494
que ou pessoa que	307

Tabela 3: Ngramas frequentes a iniciar definições de adjectivos.

Ngrama	Frequência
de modo	398
de maneira	49
muito	44
de forma	30

Tabela 4: Ngramas frequentes a iniciar definições de advérbios.

3 O analisador sintáctico PEN

O PEN é um analisador sinático, vulgo *parser*, mantido e disponibilizado gratuitamente pelo Pólo de Coimbra da Linguateca, sob uma licença BSD[1]. Este analisador encontra-se implementado na linguagem Java, de acordo com a descrição do algoritmo de Earley [5].

Trata-se de um *parser* genérico para o qual é possível construir gramáticas para os mais diversos fins (não obrigatoriamente processamento de linguagem natural), o que o torna bastante versátil.

Tem como entrada dois ficheiros de texto um com o texto a analisar e outro com a descrição de uma gramática[2]. O PEN analisa depois cada linha do texto fornecido, sendo possível visualizar a ou as derivações resultantes dessa análise.

[1] http://www.linguateca.pt/Coimbra

[2] Apesar do PEN estar preparado para receber apenas uma gramática descrita num ficheiro, essa descrição pode fazer a referência a outros ficheiros, com outras gramáticas, através de uma linha iniciada pelo caracter >, seguido do nome do ficheiro a incluir.

3.1 Formato das gramáticas

As Figuras 1 e 2 representam respectivamente o exemplo de uma gramática muito simples e a sua descrição no formato de entrada do PEN. Com base nesta gramática é possível obter derivações para algumas frases como por exemplo: "Eu li o livro" ou "Comprei um carro".

Cada regra ocupa uma linha e é identificada por uma sequência de caracteres maiúsculos. Os símbolos terminais não podem ser completamente escritos em maiúsculas, para não serem internamente "confundidos" com nomes de regras. Há ainda a ter em atenção que a primeira regra que o PEN vai verificar tem de se chamar `RAIZ`.

RAIZ → SN SV | SV
SN → DET NOME | PRON
SV → VERBO | VERBO SN
DET → o | um | ...
PRON → eu | ...
NOME → livro | carro | aviso | ...
VERBO → comprei | li | ...

Figura 1: Exemplo de gramática.

```
RAIZ ::= SN <&> SV
RAIZ ::= SV

SN ::= DET <&> NOME
SN ::= PRON

SV ::= VERBO
SV ::= VERBO <&> SN

DET ::= o
DET ::= um
(...)

PRON ::= eu
(...)

NOME ::= livro
NOME ::= carro
NOME ::= aviso
(...)

VERBO ::= comprei
VERBO ::= li
(...)
```

Figura 2: Gramática da Figura 1 no formato do PEN.

Para a frase de entrada "Eu li o aviso", obtém-se a árvore de derivação

que se pode observar na Figura 3. A representação desta árvore no formato do PEN seria semelhante à Figura 4.

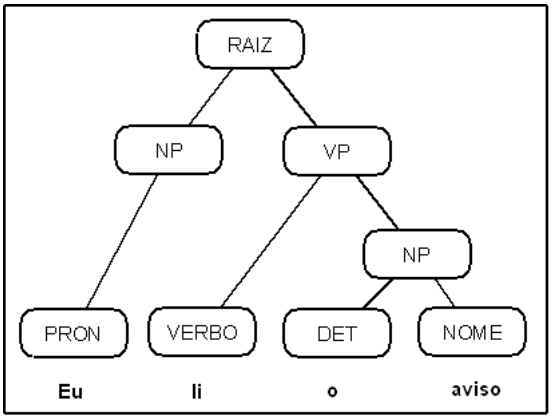

Figura 3: Árvore de derivação da frase "Eu li o aviso" de acordo com a gramática na Figura 1.

```
[RAIZ]
  [SN]
    [PRON]
      > [eu]
  [SV]
    [VERBO]
      > [li]
  [SN]
      [DET]
        > [o]
      [NOME]
        > [aviso]
```

Figura 4: Saída do PEN para a frase "Eu li o aviso".

3.2 Derivações

De acordo com a gramática, é possível que existam situações que dão origem a mais do que uma derivação para cada frase. Quando isto acontece e tendo em conta o algoritmo de Earley, todas as derivações possíveis são obtidas, ficando ao critério do programador uma eventual selecção daquelas que interessam. É também possível que simplesmente não existam derivações para uma frase, se não existir uma regra que apanhe tudo o que não é reconhecido. Para apanhar qualquer símbolo terminal não especificado utiliza-se nas gramáticas o símbolo <?> (também utilizado na gramática da Figura 5).

3.3 Peso de regras

O PEN também permite dar uma peso a cada regra da gramática, apesar de não ser um requisito. O peso pode eventualmente ser atribuído de acordo com um critério que seja depois utilizado na selecção da melhor derivação. Para atribuir um peso a uma regra basta colocá-lo em frente à regra, como na Figura 5. O peso de uma regra tem de ser um número inteiro positivo ou negativo. Não é necessário atribuir um peso a todas as regras: quando uma regra não é pontuada, o PEN assumo o valor 0 para a pontuação dessa regra.

```
RAIZ ::= SN <&> SV <&> QUALQUERCOISA
RAIZ ::= SV <&> QUALQUERCOISA

10 # SN ::= DET <&> NOME
5 # SN ::= PRON

5 # SV ::= VERBO
2 # SV ::= VERBO <&> SN

QUALQUERCOISA ::= QUALQUERCOISA <&> <?>
-1 # QUALQUERCOISA ::= <?>

...
```

Figura 5: Gramática com algumas regras pontuadas.

É possível obter o peso para cada nó independentemente ou para a sub-árvore com raíz nesse nó. O segundo será a soma do peso desse nó com os pesos de todos os nós nessa sub-árvore.

3.4 Recursividade

É muitas vezes útil utilizar regras recursivas, onde uma regra entra no seu próprio corpo.

Uma caso típico da utilização da recursividade encontra-se na Figura 5, na primeira regra do símbolo `QUALQUERCOISA`. Essa regra apanha todas as sequências de texto não especificadas.

4 Extracção de relações entre palavras

Mostrámos na secção anterior que o PEN pode ser utilizado para obter derivações de frases, de acordo com as gramáticas que lhe forem fornecidas. Compreende-se desta forma que seja possível construir gramáticas baseadas em padrões textuais para extrair informação de documentos escritos.

Como sabemos, as relações entre palavras podem ser expressas através da utilização de padrões textuais que as incluem. A obtenção e consequente utilização desses padrões em texto livre não é uma tarefa simples porque os padrões nem sempre são fáceis de identificar e têm muitas vezes uma interpretação ambígua. No entanto, se for utilizado apenas o texto das definições de um dicionário, onde o vocabulário utilizado é normalmente mais restrito e previsível [3, 4], parece ser mais simples obter esses padrões e utilizá-los em regras para extrair relações a partir das próprias definições.

4.1 Objectivo e procedimento

Para permitir a extracção de várias relações entre palavras, a partir das entradas do Dicionário da Língua Portuguesa ([1]), é necessário construir gramáticas específicas para cada uma das relações. Essas gramáticas são constituídas por regras onde estão representados padrões textuais indicadores das relações que se pretendem extrair.

De certa forma o nosso objectivo é semelhante ao que tiveram Amsler e Calzolari ([2], [3]) mas no nosso caso não nos pretendemos dedicar exclusivamente à extracção da hiperonímia, mas também à extracção de outras relações que nos pareçam plausíveis.

Utilizando o PEN é possível analisar cada entrada de um dicionário e a respectiva definição, procurando extrair relações entre a segunda e a primeira, de acordo com as gramáticas que lhe forem fornecidas, como no exemplo da Figura 6.

```
letra, s.  f.  - tipo de impressão
→ impressão HIPERONIMO_DE letra
```

Figura 6: Exemplo de relação de hiperonímia presente numa definição.

É de ter em atenção que com o processo utilizado apenas são extraídas relações entre palavras, devido a não existir qualquer sistema de desambiguação dos vários significados que uma palavra pode ter.

A relação da Figura 6 pode ser extraída com as regras da Figura 7 que têm como objectivo simples a extracção de relações de hiperonímia denotadas pelo padrão "tipo de".

```
RAIZ ::= tipo <&> de <&> HIPERONIMO_DE
RAIZ ::= tipo <&> de <&> HIPERONIMO_DE <&> QUALQUERCOISA
HIPERONIMO_DE ::= <?>

QUALQUERCOISA ::= QUALQUERCOISA <&> <?>
QUALQUERCOISA ::= <?>
```

Figura 7: Regras para a extracção de hiperonímia, utilizando o padrão "tipo de".

4.2 Obtenção da palavra relacionada

Para cada frase processada com as regras da Figura 7 ou se obtém uma derivação ou nenhuma (se por exemplo a frase não começar por "tipo de"). Nestas regras, tal como na maior parte das gramáticas criadas, foi introduzida uma regra recursiva (a que chamamos `QUALQUERCOISA`) para garantir que a derivação de uma frase é sempre completa. A derivação obtida para a definição na Figura 6 com as regras da Figura 7 encontra-se na Figura 8.

```
[RAIZ]
  > [tipo]
  > [de]
  [HIPERONIMO_DE]
    > [impressão]
```

Figura 8: Derivação de "tipo de impressão".

Como se pode verificar, existe um nó da árvore com a etiqueta HIPERONIMO_DE. Se um programa conhecer as etiquetas dos nós que foram utilizados para identificar palavras relacionadas, pode extrair relações de um tipo identificado pelo nome do nó entre a(s) palavra(s) dentro desse nó e a palavra definida (neste caso obtinha-se a relação da Figura 6 barco HIPERONIMO_DE andorinha).

4.3 Selecção da melhor derivação

No exemplo apresentado anteriormente as regras originam apenas uma derivação. No entanto isto nem sempre acontece e uma gramática pode dar origem a várias derivações, como é o caso da utilização das regras na Figura 9 para analisar a definição:

`salmonelose, s. f. - infecção causada por uma ou mais salmonelas.`

Neste caso são obtidas três derivações, representadas na Figura 10.

Se uma gramática tiver sido construída com o propósito de extrair relações apenas de um tipo, na esmagadora maioria dos casos apenas nos

```
RAIZ ::= QUALQUERCOISA <&> CAUSA_PP <&> por <&> SN
RAIZ ::= QUALQUERCOISA <&> CAUSA_PP <&> por <&> SN <&> QUALQUERCOISA

CAUSA_PP ::= causado
CAUSA_PP ::= causada

SN ::= CAUSADOR_DE
SN ::= DET <&> SN

DET ::= um
DET ::= uma
DET ::= DET <&> ou <&> mais

CAUSADOR_DE ::= <?>

QUALQUERCOISA ::= QUALQUERCOISA <&> <?>
QUALQUERCOISA ::= <?>
```

Figura 9: Regras para a extracção da relação causador, utilizando o padrão "causada por".

interessa uma derivação por gramática[3]. Para casos que dão origem a mais de uma derivação foi definido um critério para seleccionar aquela que realmente interessa: escolher aquela que tem menor número de nós do tipo `QUALQUERCOISA`. É possível verificar pela observação das três derivações na Figura 10 que um menor número de nós deste tipo significa que mais símbolos da frase foram reconhecidos pela gramática. Se isto se verificar a derivação é aquela que está mais de acordo com as regras definidas e que conterá a informação que se pretende extrair.

Generalização: Em regra geral, a palavra relacionada que se pretende obter não tem qualquer restrição (como se pode verificar nas Figuras 7 e 9 a utilização do símbolo `<?>` nas regras `HIPERONIMO_DE` e `CAUSADOR_DE`). As restrições encontram-se no padrão que atecede essa(s) palavra(s). Como as palavras relacionadas p podem ser introduzidas por palavras conhecidas c (por exemplo determinantes) é normal que, nesse caso se obtenha uma derivação de onde se extrai c (seguida de texto desconhecido) e outra em que se conhece c e se extrai p. A segunda é aquela que pretendemos e é seleccionada porque tem menos texto desconhecido (isto é, menos nós `QUALQUERCOISA`).

Por agora parece-nos que a aplicação deste critério leva a que não seja necessário utilizar pesos nas regras das gramáticas para a extracção de relações.

[3]Dificilmente uma definição inclui duas relações distintas do mesmo, obtidas através de regras diferentes. É no entanto possível obter várias relações do mesmo tipo, através da mesma regra, sendo o caso mais comum a utilização de enumerações

```
Derivação 1 (QUALQUERCOISA = 4):
[RAIZ]
  [QUALQUERCOISA]
    > [infecção]
  > [causada]
  > [por]
  [SN]
    [CAUSADOR_DE]
      > [uma]
  [QUALQUERCOISA]
    [QUALQUERCOISA]
      [QUALQUERCOISA]
        > [ou]
      > [mais]
    > [salmonelas]

Derivação 2 (QUALQUERCOISA = 3):
[RAIZ]
  [QUALQUERCOISA]
    > [infecção]
  > [causada]
  > [por]
  [SN]
    [DET]
      > [uma]
    [SN]
      [CAUSADOR_DE]
        > [ou]
  [QUALQUERCOISA]
    [QUALQUERCOISA]
      > [mais]
    > [salmonelas]

Derivação 3 (QUALQUERCOISA = 1):
[RAIZ]
  [QUALQUERCOISA]
    > [infecção]
  > [causada]
  > [por]
  [SN]
    [DET]
      [DET]
        > [uma]
      > [ou]
      > [mais]
    [SN]
      [CAUSADOR_DE]
        > [salmonelas]
```

Figura 10: Derivações para "infecção causada por uma ou mais salmonelas"

4.4 Relações directas e inversas

As relações podem estar representadas no dicionário sob a forma directa ou inversa. A Figura 11 exemplifica ambas as situações através de duas relações PARTE_DE, uma representada na forma directa e outra na forma inversa.

```
protófito, s.  m.  - planta constituída por uma única célula
                 → célula PARTE_DE protófito
punho, s.  m.  - parte da manga que cerca o pulso
             → manga INCLUI punho
```

Figura 11: Exemplos de relações PARTE_DE e a sua inversa INCLUI.

Um programa que tenha a informação de que a relação INCLUI é inversa da relação PARTE_DE pode inferir que: X PARTE_DE Y é equivalente a Y INCLUI X.

5 Agradecimentos

Este relatório foi escrito no âmbito da Linguateca, financiada pela Fundação para a Ciência e Tecnologia e pela União Europeia através dos projectos POSI/PLP/43931/2001 e POSC 339/1.3/C/NAC.

Agradecemos também a colaboração valiosa do Nuno Seco, que foi o responsável por grande parte do trabalho relatado neste relatório (extracção dos ngramas e implementação do PEN), à Diana Santos pela orientação e sugestões relativamente ao conteúdo do relatório e ainda ao Núcleo de Investigação e Desenvolvimento da Porto Editora, que nos prestou o apoio necessário do lado da Porto Editora.

Referências

[1] *Dicionário PRO da Língua Portuguesa*. Porto Editora, Porto, 2005.

[2] Robert A. Amsler. A taxonomy for English nouns and verbs. In *Proceedings of the 19th annual meeting on Association for Computational Linguistics*, pages 133–138, Morristown, NJ, USA, 1981. Association for Computational Linguistics.

[3] Nicoletta Calzolari. Detecting patterns in a lexical data base. In *Proceedings of the 22nd annual meeting on Association for Computational Linguistics*, pages 170–173, Morristown, NJ, USA, 1984. Association for Computational Linguistics.

[4] Martin S. Chodorow, Roy J. Byrd, and George E. Heidorn. Extracting semantic hierarchies from a large on-line dictionary. In *Proceedings of the 23rd annual meeting on Association for Computational Linguistics*, pages 299–304, Morristown, NJ, USA, 1985. Association for Computational Linguistics.

[5] Jay Earley. An efficient context-free parsing algorithm. *Communications of the ACM*, 6(8):451–455, 1970. Reprinted in Grosz et al. (1986).

[6] Hugo Gonçalo Oliveira, Paulo Gomes, and Diana Santos. PAPEL - trabalho relacionado e relações semânticas em recursos semelhantes, Dezembro 2007.

[7] Hugo Gonçalo Oliveira, Paulo Gomes, Diana Santos, and Nuno Seco. PAPEL - trabalho relacionado e relações semânticas em recursos semelhantes, Janeiro 2008.

Tabela de Revisões

Quem	O quê	Data
Hugo Gonçalo Oliveira	Primeira versão do relatório. Inclui partes novas e partes que tinham sido originalmente escritas para o relatório anterior	20-11-2008
Hugo Gonçalo Oliveira	Alterações sugeridas por Diana Santos, incluindo a mudança do título e a inserção de um exemplo para a selecção da melhor derivação.	25-11-2008
Hugo Gonçalo Oliveira	Mais algumas alterações sugeridas por Diana Santos.	02-12-2008
Hugo Gonçalo Oliveira	Melhor explicação do critério para seleccionar a melhor derivação.	11-12-2008