Cálculo de frequências para entradas de dicionários
através do uso conjunto de analisadores morfológicos, taggers e corpora

Paulo Alexandre Rocha
paulo.rocha@di.uminho.pt

Alberto Manuel Simões
albie@alfarrabio.di.uminho.pt

José João Almeida
jj@di.uminho.pt

Departamento de Informática
Universidade do Minho
4710-057 Braga

Resumo

Apresentamos neste documento uma possível abordagem à extracção de frequências de palavras a partir de corpora, baseada numa utilização cooperativa de várias ferramentas de Processamento de Linguagem Natural (PLN).

1 Introdução

O problema do cálculo de frequências de palavras normalizadas (lemas) para entradas de dicionários é importante em diversos domínios, nomeadamente o do ensino e aprendizagem da língua. Este problema, embora já abordado noutras línguas (Kilgarriff 1996), tem sido no entanto frequentemente negligenciado quanto ao português por ser de difícil automatização, apresentando vários problemas conceptuais. Nomeadamente, uma parte substancial das palavras gráficas (tipos) ocorrentes em textos portugueses não pode ser atribuída automaticamente a um determinado lema. Por exemplo, lava pode ser um substantivo, ou uma ocorrência do verbo lavar. Embora intuitivamente se presuma que a maior parte das ocorrências da palavra sejam relativas à forma verbal, não se pode verificar tal hipótese simplesmente contabilizando o número de existências de palavras em corpora. O significado só pode ser obtido examinando caso a caso um número significativo de ocorrências de tal palavra.

O método mais preciso para alcançar tal designio é certamente a verificação manual de cada ocorrência da palavra. Tal processo é no entanto bastante moroso. Desta forma, torna-se importante obter estes valores de uma forma automática ou, por vezes, semi-automática.

Em resumo, neste documento descreve-se o resultado da utilização conjunta de várias ferramentas, nomeadamente de:

- um analisador morfológico (jspell, e módulo Perl associado);
- um etiquetador morfo-sintáctico (EMS);
- um processador de corpora (CQP – Corpus Query Processor);
- vários corpora (CETEMPúblico, Diário do Minho, etc.).

Pretende-se obter várias estatísticas:

- taxa de ocorrências de cada entrada (valor total das ocorrências das suas flexionadas e derivadas directas).
- padrão de ocorrência dessas palavras (conjunto de variantes da palavra que realmente aparecem). Para cada um destes valores deveremos calcular uma medida de confiança (tendo em conta as ambiguidades mórfo-sintácticas).

Para que haja cooperação efectiva entre ferramentas, foi necessária uma plataforma de desenvolvimento comum: no nosso caso, optamos pela plataforma Linux, uma variante do sistema operativo Unix.

2 Ferramentas utilizadas

Nesta secção apresentamos de forma independente as várias ferramentas utilizadas.

2.1 Corpora

É naturalmente indispensável para este trabalho, uma vez que não há outro modo científico de determinar frequências.

Fazemos notar que o uso de um único corpus não balanceado pode induzir em erro o utilizador não prevenido. Por exemplo, embora nos nossos testes tenhamos usado extensivamente o corpus CETEMPúblico (Rocha & Santos, 2000), por ser um dos maiores corpora de língua portuguesa livremente disponíveis, este corpus apresenta-nos vários problemas: um dos casos mais evidentes são as formas da primeira pessoa singular que colidem com substantivos da mesma raiz (por ex.: abandono, gosto, uso) que são presumivelmente menos frequentes num texto jornalístico que num corpus que inclua, discurso falado.

2.2 Jspell

O jspell é um analisador morfológico e corrector ortográfico para português europeu, desenvolvido na Universidade do Minho (Almeida & Pinto, 1994) com base num corrector ortográfico para o inglês: ispell. O vocabulário deste programa foi recentemente amplamente expandido e corrigido utilizando diversos métodos semi-automáticos (Almeida & Simões, 2001).

A característica mais importante deste programa, para obter os resultados desejados, é indicar quais os possíveis lemas da forma flexionada de uma determinada palavra. Não pretendemos discutir aqui o significado de entrada de dicionário, cuja definição sai largamente do âmbito deste artigo. Assim, consideramos aqui as entradas cujas frequências queremos determinar como sendo as entradas separadas definidas no dicionário deste analisador morfológico.

No exemplo abaixo, verificamos qual a resposta do programa quando confrontado com uma palavra ambígua. Repara-se como as duas respostas possíveis indicam dois lemas diferentes.

```
$ jspell -d port -a
@(#) International Jspell Version 1.00b1, 11/07/2001

gosto
* gosto 0 :lex(gosto,  [CAT=nc,G=m,N=s], [], [], []),
           lex(gostar, [CAT=v,T=inf,TR=_], [], [P=1,N=s,T=p], [])
```

2.3 EMS – Etiquetador Mórfo-Sintáctico

O EMS é um etiquetador mórfo-sintáctico, ou seja, um programa que recebendo um texto devolve esse mesmo texto com a respectiva anotação gramatical. Aqui interessa-nos principalmente a categoria gramatical da palavra – outras características como género e número são menos relevantes para o nosso trabalho, embora não completamente inúteis. Abaixo apresentamos um exemplo de frase etiquetada com o EMS:

```
Governo/NCMS impõe/VIH3S limites/NCMP durante/P seis/DNCNP meses/NCMP
```

Cada uma das etiquetas que aparece à direita da palavra caracteriza-a utilizando a seguinte convenção:

NC nome comum;

VI verbo intransitivo;

P preposição;

DNC determinante numeral cardinal.

Para casos de palavras ambíguas, o EMS começa por atribuir a cada palavra uma categoria (baseada numa das respostas do jspell), e em seguida, de acordo com o seu contexto, (classificação gramatical das palavras vizinhas), tenta corrigir a classificação inicial.

Por exemplo, a frase "eu gosto do gosto da batata" começa por ser etiquetada incorrectamente:

```
Eu/PSN1S gosto/NCMS do/\&MS gosto/NCMS da/\&FS batata/NCFS ./.
```

Note-se como o primeiro "gosto" é classificado como nome comum. No entanto, esta etiquetação inicial é modificada através de uma regra de modificação existente num ficheiro de regras

```
NCMS VIH1S PREVTAG PSN1S
```

Ou seja, nomes comuns masculinos singulares (NCMS) são transformados em primeiras pessoas do singular do presente do indicativo (VIH1S), no caso de a etiqueta anterior (PREVTAG) ser o pronome pessoal da primeira pessoa do singular (PSN1S).

O resultado final, neste caso, é uma frase correctamente etiquetada.

```
$echo "eu gosto do gosto da batata."| ems

eu/PSN1S gosto/VIH1S do/\&MS gosto/NCMS da/\&FS batata/NCFS ./.
```

2.4 Módulo CQP

O Corpus Query Processor (CQP) é parte do Corpus Workbench do Institut für Maschinelle Sprachverarbeitung (IMS), da Universidade de Estugarda (Christ et al. 1999). Este programa corre em ambiente Linux e tem tido um comportamento satisfatório mesmo quando usado com corpora de grandes dimensões. Apresentamos abaixo alguns exemplos do uso desta ferramenta para a extracção de concordâncias:

```
CETEMPUBLICO> "Putin";
 43588961: eira confiança , Vladimir <Putin> , que ocupava o cargo de
113251175: a substituição , Vladimir <Putin> , primeiro adjunto do ch
154049540: urança ( FSB ) , Vladimir <Putin> , e com o chefe do servi
```

O CQP permite ainda a utilização de corpora anotado, como se pode ver no exemplo abaixo, e nada objecta ao uso directo de um corpora anotado morfo-sintacticamente. Na verdade, o projecto Processamento Computacional do Português disponibiliza a consulta na Rede de vários corpora gramaticalmente anotados. (`http://corpora.portugues.mct.pt/anotado.html`).

O exemplo abaixo apresenta um extracto retirado de um destes corpora.

```
CETEMPANOT> "gosto";
     7126: PEC_rel me/PERS_refl dá/V <gosto/N> imaginar/V como/ADV_inte
    95008: essiva/ADJ de/PRP mau/ADJ <gosto/N> ./PU No/PRP+DET_artd pró
   135739: ão/N que/SPEC_rel não/ADV <gosto/V> de/PRP beringelas/N amei
   144462: U como/ADV_rel_ks eu/PERS <gosto/V> de/PRP dizer/V ,/PU um/D
```

Para facilitar o uso sistemático e repetitivo deste programa, foi criado um módulo Perl (CQP.pm). Embora se pudesse utilizar directamente o CQP, a construção deste módulo permitiu um grande número de facilidades que de outra forma não seriam integráveis tão facilmente na programação em Perl. Em particular, uma das funções, que dada uma palavra e um valor inteiro n retorna uma lista com n frases onde ocorre essa palavra foi bastante útil.

3 Esquema utilizado

A parte mais simples é obviamente a extracção de todos os tipos (ou seja formas gráficas distintas) e respectivas frequências existentes num corpora ou conjunto de corpora. O exemplo abaixo mostra um extracto de uma destas listas de frequência:

```
  6100  gosta
 11835  gosto
  3207  gostava
  2551  gostou
  1718  gostos
```

Algumas destas formas não são ambíguas (gostou, gostos), e as suas ocorrências podem ser automaticamente atribuídas à palavra base correspondente (respectivamente, o verbo gostar e o substantivo gosto).

No entanto, gosto é uma palavra ambígua na língua escrita (embora não o seja na língua falada), uma vez que além de um substantivo, pode ser igualmente uma forma de verbo gostar, e só a análise do contexto nos permitirá determinar em que proporção as ocorrências desta palavra gráfica se repartem pelas duas palavras base.

Desta forma, depois de calculadas as ocorrências de cada palavra e detectadas as ambiguidades podemos já criar uma lista com a respectiva confiança. Segue-se um pequeno algoritmo para obter estes resultados:

```
use jspell;
jspell_dict("port");

foreach (numoco, palavra) in stdin
     w = lemas(palavra)
     duvida = (comprimento(w) > 1);

     foreach (lema) in (w)
         if (duvida) { oco[p][duv] += numoco; }
         else            { oco[p][gar] += numoco; }

foreach (palavra) in (sort dom(oco))
    total = oco[palavra][duv] + oco[palavra][gar];
    conf = oco[palavra][gar]/total;
    print palavra, total, conf;
```

Depois de executar este programa, obtemos uma lista como a que se segue:

```
abade       571 (conf=100%)
abadia      270 (conf=100%)
abafar     1103 (conf=98%)
...
gostar    43194 (conf=70%)
...
gosto     14400 (conf=12%)
```

Esta ferramenta permite obter resultados mais complexos bastando para isso alterar as opções de linha de comando. Uma das possibilidades é a geração de um ficheiro Perl que pode ser incluído directamente em qualquer script Perl. Isto permite que se perca muito menos tempo no desenvolvimento das aplicações.

Outra opção, legível pelo utilizador comum, pode ser obtida com uma flag denominada `-complex`, que retorna uma lista no seguinte formato:

```
ajuda (:229)
      ajuda(176) ajuda/ajudar
      ajudas(53) ajuda/ajudar
ajudar (175:229)
      ajuda(176) ajuda/ajudar
      ajudar(140)
      ajudas(53) ajuda/ajudar
      ajudam(15)
      ajudaram(10)
      ajude(10)
```

Embora nos tenhamos apoiado em ferramentas como o etiquetador morfo-sintáctico desambiguação, poderíamos ter usado métodos semi-estatísticos para resolver este problema. Uma das soluções seria o cálculo médio de ocorrências de verbos e de substantivos, realizando posteriormente a divisão das frequências baseado nestes valores. É sem dúvida uma solução possível mas, uma vez que temos acesso a outras ferramentas, torna-se obrigatório o uso de métodos mais civilizados.

Depois do cálculo da taxa de ocorrências e da certeza respectiva podemos refinar alguns destes valores através da desambiguação das ocorrências de formas gráficas ambíguas extraindo frases exemplo da palavra em causa (usando o módulo CQP.pm) e classificando morfo-sintacticamente o extracto com base no contexto gramatical (usando o EMS).

```
A/DADFS Direcção/NCFS e/C o/DADMS treinador/JMS querem/VIH3P que/QUE
fique/VSH_S e/C eu/PSN1S gosto/VIH1S do/\\&MS Beira/NPMS Mar/NPMS ,/,
por/P isso/PDNN .../...

No/\&MS fundo/NCMS ,/, transformando/VG Portugal/NPNN num/\&MS país/NCMS
no/\&MS qual/PRANS dê/VSH1S gosto/NCMS viver/VN ./.

À/ADV passagem/NCFS do/\&MS último/JMS quarto/DNOMS de/P hora/NCFS ,/,
de/P novo/JMS João/NPMS Paulo/NPMS voltou/VIP3S a/P fazer/VN o/DADMS
«/« gosto/NCMS ao/\&MS pé/NCMS »/» ,/, aproveitando/VG um/DAIMS bom/JMS
passe/NCMS de/P Carlitos/NPMS ./.

Quanto/ADV ao/\&MS demais/JNS ,/, tive/VIP1S o/DADMS gosto/NCMS
pessoal/JNS de/P saber/NCMS que/QUE ,/, pelo/\&MS menos/ADV ,/,
quatro/DNCNP distritos/NCMP do/\&MS país/NCMS ,/, gostariam/VCH3P que/QUE
fosse/VSI3S deputado/JMS por/P esses/PDMP círculos/NCMP ./.

Mas/C ,/, gosto/NCMS de/P lidar/VN com/P as/DADFP coisas/NCFP na/\&FS
base/NCFS da/\&FS verdade/NCFS ./.

Prova/NCFS disso/\&NN mesmo/JMS é/VIH3S o/DADMS facto/NCMS de/P todo/PFMS
este/PDMS trabalho/NCMS ser/VN feito/JMS apenas/ADV por/P gosto/NCMS ./.
```

Estes resultados permitem que se contem o número de ocorrências de determinada categoria nos extractos realizados. Do número obtido, podemos calcular uma percentagem da probabilidade de uma palavra pertencer a essa categoria. Desta forma, multiplica-se o número de ocorrências ambíguas por este valor obtendo um valor provável para o número de ocorrências. Como se trata apenas de um extracto, não tem necessariamente de corresponder à verdade, mas o que desejamos, na verdade, é apenas de uma estimativa. Além deste problema de estarmos a lidar com um extracto, temos o problema de alguma má classificação da palavra, como se pode ver na quinta frase do exemplo anterior.

Usando o etiquetador PALAVRAS (Bick 2000), das 7.338 ocorrências do tipo gosto, 5.144 foram lematizadas com o substantivo, enquanto 2.187 forma lematizadas com o verbo gostar. No

entanto, não há qualquer obrigatoriedade de analisar uma quantidade tão grande de ocorrências – amostras mais pequenas, desde que provenham de um corpus balanceado, podem entregar igualmente resultados válidos.

Realizado este processo sistematicamente para cada uma destas palavras ambíguas – o que pode ser um processo demorado, dada a grande quantidade de palavras a examinar; obtemos assim finalmente a lista ordenada de ocorrências de entradas de dicionários pretendida.

4 Uso de corpora anotado

O uso de corpora anotado permite evitar muitos problemas, partindo do princípio que a anotação do corpus está (maioritariamente) correcta. Por exemplo, estando cada palavra anotada com o devido lema, podemos usar uma ferramenta incluída no IMS-CWB, o lexdecode, para obter uma lista da quantidade de ocorrências de cada lema, como no exemplo abaixo, em que usamos o corpus etiquetado com o PALAVRAS.

```
$ lexdecode -f -P lema -p abandon.* cetempanot
      17507 abandonar
       4180 abandono
         68 abandonar+se
```

Note-se que, neste caso, a maior parte das ocorrências do lema abandonar são em casos onde não há ambiguidade quanto ao lema (abandone, abandonassemos, abandonando, etc.).

Usando este método, embora os resultados finais sejam extremamente rápidos de obter, assume que o etiquetador usado no corpus é estável e fiável, uma vez que o processo de anotação do corpus, com os etiquetadores actualmente disponíveis, é extremamente demorado. Além disso, não permite a distinção entre palavras com o mesmo lema. Alguns lemas ambíguos podem ser no entanto distinguidos com recurso a uma análise morfológica. Por exemplo, existem pelo menos três significados possíveis não figurativos do lema "lama". Um desses significados (lodo) pode ser frequentemente identificado automaticamente devido às suas características morfológicas (nomeadamente, é um substantivo feminino); no entanto, a distinção entre os dois lemas que são substantivos masculinos (sacerdote budista e mamífero sul-americano) exige uma análise manual.

5 Cálculo de entradas em dicionário

A tradição de entradas de dicionários leva a que dividam as ocorrências por patamares (ou degraus) que podem ser definidos de várias formas, dada a lista de lemas ordenadas por ocorrências

- dividi-los em grupos contendo igual número de lemas (por exemplo, os mil lemas mais frequentes, os mil lemas seguintes, etc.);
- dividi-los em k grupos, cada um correspondendo a 1/k das frequências do corpus;
- dividir de acordo com o número de ocorrências usando uma escala logarítmica.

```
$ lexdecode''f CETEMPANOT > ocorrencias
$ freqnormpt''oco''steps=1000:2000:3000 ocorrencias
  ...
  de      3
  a       3
  o       3
  que     3
  ...
  desportivo 2
  ferir      2
  curto      2
```

```
...
colaborador 1
erguer      1
mercadoria  1
...
mobiliário  0
acender     0
suspeitar   0
...
```

Neste exemplo usamos a primeira opção, em que os números indicados correspondem à identificação do patamar em que a palavra se encontra.

6 Problemas

Um problema já referido é referente aos casos de corpora não balanceados. Por exemplo, na tabela seguinte, mostramos as ocorrências por milhão de palavras de alguns lemas em dois corpora que, embora ambos baseados em texto jornalístico, apresentam um uso de vocabulário extremamente diverso.

Avante	Diário do Minho
camarada(s)	arcebispo(s)
170,64,3	2,478,0

Assim sendo, para o bom desenrolar deste projecto é essencial usar um conjunto de corpora de textos diversificado e bem balanceado. Reunindo corpora de diferentes fontes, pode-se inclusive excluir das listas palavras cuja elevada ocorrência no corpus se deve a uma única fonte (Kilgarrif 1996).

Em relação a este problema foi desenvolvida uma pequena ferramenta para permitir a comparação entre a frequência de palavras entre corpora distinto. Neste caso, é possível obter dois grupos de palavras, umas mais frequentes num dos corpora e o outro, mais frequente no outro corpus. Esta possibilidade permite o desenvolvimento de ferramentas de classificação automática de assuntos baseando-se na anterior aprendizagem sobre textos pré-classificados.

Este método não dá naturalmente nenhuns resultados quanto a palavras homógrafas pertencentes à mesma categoria gramatical (saber, banco, etc.); embora o princípio do uso de corpora continue a ser válido, é indispensável para estes casos uma cuidadosa verificação manual, caso se pretenda efectivamente incluir no dicionário tal distinção; uma análise automática é possível (Stevenson & Wilks, 2001), mas é bastante trabalhosa e o grau de fiabilidade é substancialmente baixo.

7 Conclusões

Embora saia do âmbito deste documento uma avaliação aos diversos etiquetadores existentes, os resultados do EMS não são ainda completamente satisfatórios, o que impedem uma análise mais correcta dos resultados finais. Por outro lado, com o objectivo de melhorar o mesmo etiquetador não é vantajosa a etiquetação sistemática de todo o corpus devido ao seu grande tamanho.

Uma conclusão óbvia é que é extremamente vantajoso utilizar directamente corpora anotados. No entanto, como dito acima, o processo de anotar grandes corpora é demorado, e só deve ser usado com etiquetadores em fase estável do desenvolvimento.

7.1 Trabalho futuro

Pretendemos melhorar a interface CQP.pm, que ainda está numa fase experimental e ligeiramente ineficiente. Continuam também os trabalhos de melhoria do etiquetador morfo-sintáctico.

Também deveríamos testar outros etiquetadores e realizar contagens sobre corpora já anotados, realizadas extensivamente para que se pudessem comparar os resultados obtidos com contagens realizadas por outros grupos (p.ex. o Léxico Multifuncional Computorizado do Português Contemporâneo).

Referências

IMS Corpus Workbench http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/

Léxico Multi-funcional Computorizado do Português Contemporâneo http://www.clul.ul.pt/sectores/projecto_lmcpc.html

Projecto Processamento Computacional do Português http://www.portugues.mct.pt/

Projecto Natura http://natura.di.uminho.pt onde as várias aplicações estão disponíveis.

Almeida, J.J. e Ulisses Pinto , "Jspell" um módulo para análise léxica genérica de linguagem natural", Actas do Congresso da Associação Portuguesa de Linguística, Évora, 1994. http://www.di.uminho.pt/~jj/pln/jspell1.ps.gz

Bick, Eckhard . The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Aarhus University Press, 2000.

Christ, Oliver, Bruno M. Schulze, Anja Hofmann & Esther König . The IMS Corpus Workbench: Corpus Query Processor (CQP): User's Manual. Institute for Natural Language Processing, University of Stutgart, March 8, 1999 (CQP V2.2), http://www.ims.uni_stuttgart.de/projekte/CorpusWorkbench/CQPUserManual/HTML/

Kilgarriff, Adam . "Putting Frequencies in the Dictionary"In International Journal of Lexicography 10 (2), 1997, pp. 135-155 ftp://ftp.itri.bton.ac.uk/reports/ITRI-96-10.ps.gz

Rocha, Paulo Alexandre & Diana Santos . "CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa", Actas do V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR'2000) (Atibaia, São Paulo, Brasil, 19 a 22 de Novembro de 2000), pp.131-140. http://www.portugues.mct.pt/Diana/download/RochaSantosPROPOR2000.pdf

Santos, Diana & Eckhard Bick . "Providing Internet access to Portuguese corpora: the AC/DC project", in Maria Gavrilidou et al. (eds.), Proceedings of the Second International Conference on Language Resources and Evaluation, LREC 2000 (Athens, 31 May-2 June 2000) http://www.portugues.mct.pt/Diana/download/SantosBickLREC2000.rtf

Stevenson, Mark & Yorick Wills . 2001. The Interaction of Knowledge Sources in Word Sense Disambiguation. In Computational Linguistics, v.27, n.3, pp. 321-350.