Extracção de Definições no Corpógrafo
Relatório

Ana Sofia Pinto
Débora Oliveira
Pólo FLUP da Linguateca
Faculdade de Letras da Universidade do Porto
anasofianovaispinto@yahoo.com
deboramso@yahoo.com

Outubro 2004

Introdução

O presente relatório tem como objectivo dar conta das actividades desenvolvidas no pólo FLUP da Linguateca no âmbito da funcionalidade de extracção de definições da ferramenta Corpógrafo.

O Corpógrafo[1] é uma plataforma destinada à análise e ao trabalho sobre corpora com o objectivo essencial de produzir recursos linguísticos vários em suporte informático. O Corpógrafo pretende apoiar os investigadores da língua portuguesa num conjunto de tarefas que vão desde a compilação de corpora, à extracção e organização do conhecimento gerado a partir deles.

Uma das funcionalidades desta ferramenta é a extracção semi-automática de definições. Esta é uma área de extracção semi-automática de definições é uma área ainda pouco explorada não havendo, por isso, grande quantidade de informação disponível relativamente aos passos a realizar em cada fase do processo de extracção (identificação, compilação, teste, validação e consolidação de padrões) de definições.

Sendo assim, o trabalho realizado pelo pólo FLUP nesta área teve de ser desenvolvido de raiz. Seguidamente, descrevemos em detalhe esse trabalho.

Definição

Consideramos definição toda a frase ou segmento de frase que possa descrever/caracterizar um dado termo.

Exemplo 1: *"Os neurónios são células muito especializadas que apresentam um ou mais prolongamentos, ao longo dos quais se desloca um sinal eléctrico."*

Exemplo 2: *"Cada neurónio compreende um corpo celular (…) e um prolongamento, o axónio, que pode ser muito longo e apresentar ramificações na sua parte distal ou , ao longo da sua extensão, formando ramificações colaterais."*

Exemplo 3: *"A Fagocitose ocorre apenas em células especializadas, como os macrófagos e os granulócitos, e envolve a digestão de partículas de grandes dimensões (ex.: vírus , bactérias , detritos celulares)."*

[1] O Corpógrafo é financiado pela Fundação para a Ciência e Tecnologia, co-financiada pelo POSI, através do projecto POSI/PLP/43931/2001 (Linguateca).

Método

Todo o trabalho foi desenvolvido tendo por base um corpus de 172.482 palavras subordinado ao tema “neurónios”. O corpus, de nome “Neurodemo”, é constituído por textos de cariz científico em português europeu, português do Brasil, inglês, francês, alemão, italiano e espanhol.

Depois de todos os ficheiros terem sido pré-processados, foi feita a pesquisa terminológica, tendo sido extraído um número bastante elevado de termos para cada uma das línguas (ver Tabela 1).

Língua	Termos Extraídos
Português Europeu (PT_PT)	270
Português do Brasil (PT_BR)	144
Inglês (EN)	661
Alemão (DE)	109
Italiano (IT)	215
Francês (FR)	203
Espanhol (ES)	190

Tabela 1 - Número de termos extraídos do corpus "Neurodemo" para cada língua.

Perante a inexistência de uma lista de padrões para extracção de definições já compilada, foi necessário proceder à pesquisa e indexação dos mesmos. Por ser um processo moroso, apenas se pesquisaram os padrões para inglês e português europeu (54.340 palavras). O processo de pesquisa foi desenvolvido em várias fases.

A primeira fase consistiu na pesquisa e análise das colocações de cada um dos termos existentes na base terminológica no sentido de identificar quais as estruturas sintácticas que fazem com que uma frase, ou segmento de frase, seja considerada uma definição.

Exemplo: *”O neurónio é uma célula ramificada, complexa e altamente diferenciada, com uma estrutura básica comum a todas as partes do sistema nervoso central (…).”*

Após a identificação dessas estruturas, estas foram transliteradas em padrões utilizando expressões regulares Perl. Estes padrões foram compilados em ficheiros diferenciados com base no número (Singular e Plural).

Exemplo 1:

“O neurónio é uma (…).” = (^|^o |^a |^um |^uma | o | a | um | uma)__TERMO__ é (o|a|um|uma)

Exemplo 2:

“Os neurónios são (…).” = (^|^os |^as | os | as)__TERMO__ são

À medida que os padrões iam sendo compilados, estes iam sendo testados no corpus “Neurodemo”, de forma a verificar se os padrões estavam correctamente formulados.

Depois de compilados, todos os padrões existentes neste corpus foram também testados no corpus "Busca". Este corpus é constituído por 56 textos em português europeu (439.215 palavras) e 107 textos em inglês (441.990 palavras) nas áreas do Processamento da Linguagem Natural e da Linguística. Os testes realizados neste corpus permitiram não só validar os padrões já existentes, como também aumentar as listas de padrões anteriormente compiladas.

Sabendo que na lista de padrões para o Singular existiam padrões que poderiam ser pluralizados e vice-versa, estes foram identificados e, seguidamente, adicionados às respectivas listas.

Exemplo 1:
Singular – (^|^o |^a| o | a)__TERMO__ é transformad(o|a) em
"**A informação sensorial é transformada em** sinais eléctricos pelas células receptoras ou sensorais, …"
Plural – (^|^os |^as | os | as)__TERMO__ são transformad(os|as) em

Exemplo 2:
Plural – (^|^os |^as | os | as)__TERMO__ são produzid(os|as) (por|na|no||nas|nos)
"…**as hormonas são produzidas por** células especializadas…"
Singular – (^|^o |^a | o | a)__TERMO__ é produzid(o|a) (por|no|na|nos|nas)

Posteriormente, pôs-se a hipótese de existirem padrões em português europeu que pudessem também funcionar para extracção de definições em inglês e vice-versa (tanto no singular, como no plural). Esses padrões foram identificados, traduzidos para as respectivas línguas e acrescentados às listas.

Exemplo 1:
PT_PT – (^|^o |^a | o | a)__TERMO__ que é (o|a|um|uma)
"…**núcleo celular , que é a** central de energia da célula ."
EN – (^|^the |^a |^an | the | a | an)__TERMO__ (that|which) is (the|a|an)

Seguidamente, procedeu-se à consolidação dos padrões, que consistiu na identificação dos padrões que poderiam ser comprimidos num só, fazendo com que este se tornasse mais abrangente.

Exemplo 1: (^|^o |^a |^um |^uma | o | a | um | uma)__TERMO__ é (constituíd|determinad|compost)(o|a) (por|pelo|pela|pelos|pelas|de)

Exemplo 2: (^|^os |^as | os | as)__TERMO__ (originam-se|designam-se|ligam-se|dão-se|estendem-se)

Tendo em conta que estes padrões foram extraídos de um corpus de um domínio específico, nomeadamente o da Neuroanatomia, sentiu-se a necessidade de testá-los num corpus de uma outra

área de conhecimento (corpus de referência) para verificar se estes seriam eficazes num outro domínio.

Optou-se, na área da Medicina, pelo domínio específico da Fibromialgia. Depois de alguma pesquisa, verificou-se que este é um domínio com bastante terminologia específica e que permite encontrar facilmente documentos ricos em definições, tornando-se por isso uma óptima matéria-prima para testar e validar os padrões. Compilou-se, assim, um corpus com 10 textos em português europeu (21.667 palavras) e 23 textos em inglês (80.295 palavras).

Língua	Número de Termos
Alemão	140
Inglês	816
Português PT	161

Tabela 2 - Termos extraídos do corpus "Fibromialgia" para cada língua.

Seleccionaram-se 30 termos em inglês e 30 em português com base, fundamentalmente, na frequência de ocorrências e também na relevância para o domínio específico.

Resultados

Os resultados dos testes realizados no corpus de referência são os seguintes:

Termo	Nº de Definições Existentes	Nº de Definições correctamente Extraídas pelo Corpógrafo	Nº de Definições Propostas pelo Corpógrafo
autonomic nervous system	2	1(50%)	1 (100%)
central nervous system	1	0 (0%)	0 (0%)
chronic fatigue syndrome	1	0(0%)	0 (0%)
chronic pain	1	0(0%)	0 (0%)
craniosacral system	1	0(0%)	0 (0%)
craniosacral therapy	1	1 (100%)	1 (100%)
fibromyalgia	31	29 (94%)	33 (88%)
fibromyalgia diagnosis	1	0(0%)	0 (0%)
fibromyalgia pain	2	1 (50%)	1(100%)
fibromyalgia syndrome	3	3(100%)	4 (75%)
irritable bowel syndrome	1	1(100%)	4 (25%)
muscle relaxant	1	1(100%)	0 (0%)
muscle twitch	1	0(0%)	0 (0%)
myofascial pain	1	1(100%)	2 (50%)
myofascial pain syndrome	2	2(100%)	2(100%)
nocioceptive system	1	0(0%)	0 (0%)
pressure points	1	0(0%)	0 (0%)
primary fibromyalgia	2	1 (50%)	1(100%)
primary fibromyalgia syndrome	1	0(0%)	0 (0%)
restless leg syndrome	1	1(100%)	1(100%)
rheumatoid arthritis	1	1(100%)	1(100%)
secondary fibromyalgia	3	2 (67%)	2(100%)
sleep disorder	1	0 (0%)	0 (0%)
substance P	4	1 (25%)	1(100%)
sympathetic nervous system	3	0(0%)	0 (0%)
tender point	6	2 (33%)	2(100%)
tricyclic antidepressant	1	0 (0%)	0 (0%)
trigger point	4	4(100%)	4(100%)
trigger point therapy	2	1 (50%)	1(100%)
trigger point pain	1	1(100%)	1(100%)

Tabela 3 - Resultados dos testes realizados no corpus "Fibromialgia" para inglês.

Term	Existing Definitions	Definitions correctly extracted with the Corpógrafo	Definitions Proposed by the Corpógrafo
amitriptilina	1	1(100%)	1 (100%)
biofeedback	1	0(0%)	0 (0%)
ciclobenzaprina	1	1(100%)	1 (100%)
clonazepam	1	1(100%)	1 (100%)
disfunção miofascial	1	0(0%)	0 (0%)
dor crónica	2	1 (50%)	1 (100%)
dor da fibromialgia	1	1(100%)	2 (50%)
dor psicogénica	1	0(0%)	0 (0%)
esqueleto axial	1	0(0%)	0 (0%)
fadiga crónica	3	3 (100%)	3 (100%)
fáscia	6	5 (100%)	5 (100%)
fibromialgia	50	16 (32%)	23 (70%)
libertação mio fascial	6	5 (83%)	5 (100%)
líquido céfalo raquidiano	1	0(0%)	0 (0%)
modelo biopsicossocial	1	0(0%)	0 (0%)
orfenadrina	1	1(100%)	1 (100%)
personalidade pró-dolorosa	1	1(100%)	1 (100%)
ponto doloroso	2	2(100%)	2 (100%)
rigidez matinal	1	1(100%)	1 (100%)
serotonina	1	0(0%)	0 (0%)
SFM	1	1(100%)	1 (100%)
síndrome de fadiga crónica	3	3(100%)	4 (75%)
síndrome de fibromialgia	1	0(0%)	0 (0%)
síndrome do cólon irritável	1	0(0%)	0 (0%)
sistema sacro craniano	3	0(0%)	0 (0%)
substância P	2	0(0%)	0 (0%)
suplemento natural	2	2(100%)	2 (100%)
tecido conectivo	1	1(100%)	1 (100%)
terapia cognitivo-comportamental	2	1 (50%)	1 (100%)
terapia sacro-craniana	3	2 (67%)	2 (100%)

Tabela 4 - Resultados dos testes realizados no corpus "Fibromialgia" para português europeu.

Conclusões

As listas de padrões compiladas não são estanques. Tendo em conta que estes padrões foram extraídos de um corpus de texto corrido e não de um corpus de glossários ou manuais, não podemos esperar que estas listas compreendam todos os padrões existentes para extracção de definições. Assim sendo, estas listas estarão sempre incompletas e, ao mesmo tempo, em crescimento.

Além disso, há padrões que, devido à sua abrangência, trazem por vezes algum ruído. No entanto, consideramos que a percentagem de ruído é, em média, bastante inferior à percentagem de definições que estes padrões conseguem extrair.

Exemplo: (^|^o |^a |^um |^uma | o | a | um | uma)__TERMO__ é
Nas 72 ocorrências deste padrão, 62 são definições, facto que justifica a manutenção deste padrão.

Para além disso, o factor determinante para a percentagem de imprecisão obtida é o corpus com que se trabalha. Consequentemente, optámos por manter esses padrões nas listas.

Concluímos também que os padrões apresentam já uma percentagem muito razoável de eficácia. Verificámos que os padrões permitem extrair um número considerável de definições.

Finalmente, verificámos que, apesar de se encontrar ainda em desenvolvimento, esta funcionalidade do Corpógrafo mostra-se já muito eficaz nos seus resultados, sendo por isso útil para os seus utilizadores.

Futuramente

Como já foi dito, este é um trabalho em desenvolvimento e, como tal, esta funcionalidade continuará a ser optimizada. Essa optimização passará pela extracção, consolidação e validação de padrões para alemão, italiano, francês e espanhol, assim como pela adição de novos padrões para inglês e português.

Adicionalmente, todos os padrões serão revistos, tentanto agrupar as opções neles existentes através de categorias de alternativas possíveis criadas para o efeito. Desta forma, será possível reduzi-los a uma expressão muito mais compacta, melhorando a sua leitura e permanente manutenção.

Exemplo 1: **(^|^o |^a |^um |^uma | o | a | um | uma)**__TERMO__ é

Nos padrões que tenham este prefixo (a negrito), este será substituído pela etiqueta **Prefixo de Classe 1**, ou seja:

Prefixo de Classe 1__TERMO__ é

Exemplo 2: __TERMO__ **(origina|desencadeia|causa|provoca)**

Nos padrões que tenham este sufixo (a negrito), este será substituído pela etiqueta **Sufixo Classe 1**, ou seja:

__TERMO__ **Sufixo Classe 1**

Além disso, como a pesquisa de padrões é feita sequencialmente, há um conjunto de optimizações que se prende com a ordem segundo a qual os padrões serão pesquisados. Assim, será feito um estudo que fundamentará uma posterior ordem dos padrões, do mais restrito para o mais genérico, aumentando consequentemente a precisão global do processo.

Exemplo 1:
(^|^o |^a |^um |^uma | o | a | um | uma)__TERMO__ é (constituíd|determinad|compost)(o|a) (por|pelo|pela|pelos|pelas|de) – *Padrão restrito*

(^|^o |^a |^um |^uma | o | a | um | uma)__TERMO__ é – *Padrão genérico*

A metodologia usada na identificação e validação de padrões para extracção de definições está actualmente a ser aplicada na identificação de padrões para extracção de relações semânticas, nomeadamente merónimos e hipónimos. Esperamos vir a ter resultados tão bons na extracção de relações semânticas como os da extracção de definições.

Referências

1. Greenwood, M. A. & Saggion, H. *A Pattern Based Approach to Answering Factoid, List and Definition Questions*. In *Proceedings of the 7th RIAO Conference (RIAO 2004)*. Avignon. France. April 27, 2004.
2. Klavans, J. L. & Muresan, S. *Evaluation of DEFINDER: A System to Mine Definitions from Consumer-oriented Medical Text*. In *Proceedings of JCDL*. 2001.
3. Klavans, J. L. & Muresan, S. *Evaluation of the DEFINDER System for Fully Automatic Glossary Construction"*. In *Proceedings of AMIA*. 2001.
4. Morin, E. & Martienne, E. *Using a Machine Learning Tool to Refine Patterns. Actes, 11th European Conference on Machine Learning (ECML'00)*. Barcelona, Spain. 2000.
5. Muresan, S., Popper, S. D., Davis, P. T. & Klavans, J. L. *Building a Terminological Database from Heterogeneous Definitional Sources*. In *Proceedings of the National Conference on Digital Government Research*. Boston, Massachusetts. 2003.
6. Plamondon, L. & Kosseim, L. *QUANTUM: A Function-Based Question Answering System*. In *Proceedings of the Fifteenth Canadian Conference on Artificial Intelligence (AI'2002)*. R. Cohen & B. Spencer (Eds.). *Lecture Notes in Artificial Intelligence no. 2338*, pp 281-292. Springer-Verlag. Berlin. May 2002. Calgary, Canada.
7. Pompidor, P., Sala, M. & Hérin, D. *Within the Framework of course-assisted Creation, an Incremental Method to Extract Relevant Information from the Web and Integrate it in a Course Draft*. In *Proceedings of the International Workshop on Semantic Web for Web-based Learning (SW-WL'03) – Implications in the area of Educational Information Systems, connected with the 15th International Conference on Advanced Information Systems Engineering (CAISE)*. Klagenfurt/Velden, Austria, June 2003, pp 265-274.
8. Watrin, Patrick. *Information Extraction and Lexicon-Grammar*. In *Proceedings of the Dutch-Belgian Information Retrieval Workshop (DIR 2003)*. Amsterdam. 2003.