# Extracção Automática de Terminologia

## Uma breve abordagem

Luís Sarmento

Simpósio Doutoral da Linguateca

Lisboa, 6 de Maio de 2005

# Resumo

- O que é Terminologia (11)
- Para que serve a Terminolgia (3)
- Como Extrair Terminolgia (14)
  - Vários métodos
- Um caso particular (14)
  - Corpógrafo
- Conclusões (2)

# Motivação para a EAT

- Os “termos” são uma parte essencial da comunicação de domínio específico
- Identificação é fundamental para:
  - Análise / Compreensão
  - Geração
  - Tradução
- Constante aumento da informação disponível de domínio específico (DE)
- A Extracção Automática de Terminologia (EAT) torna-se fundamental (iremos já ver exemplos)

# “Termos”

- O que são os “termos”?
- Primeiro vamos alterar a nossa própria terminologia
  - “termo” → Unidade Terminológica : UT
- Evitar confusões com a terminologia usada em IR
  - “termo” → 1 qq palavra (para indexação)

# Unidades Terminológicas

- O que é uma UT?
  - Questão complexa e controversa
- Mas poderemos dizer:
  - Uma UT é uma etiqueta linguística para um conceito:
    - Domínio → Conceito – UT
- Visão clássica do Conhecimento:
  - Estruturado, dividido em domínio, redes conceptuais
- Na prática as estrutura de conhecimento são
  - dinâmicas, dependentes da perspectiva e da utilização
- Mas podemos ficar com esta ideia:
  - UT como uma "representação lexical" de um conceito

# Unidades Terminológicas

- UT dividem-se em:
  - Simples / 1 palavra
    - “rede”, ”célula”, “sistema”
    - Normalmente genéricas e ambíguas, até em domínios fechados
  - Compostas / multi-palavra
    - “rede neuronal” / “célula eucariótica”, “sistema de bombagem de água”
    - Normalmente não ambíguas num determinado domínio, mas admitem muitas variações

# Unidades Terminológicas

- Mais terminologia sobre UT's.
- UT's são habitualmente formadas:
  - por uma *cabeça*
    - termo muito genérico: "rede", "célula","ácido"
    - termo secundário: "sistema", "processo", "estrutura"
  - Um *grupo modificador* ou *grupo de argumentos*
    - Adjectivo, nome ou **outra UT**
      - [Sistema de [gestão de [bases de dados]]]
  - Certas UT formam-se por aglutinação mas em português são raras ou só ocorrem em certos domínios (química)

# Quanto à Morfologia

- É difícil descrever a morfologia dos termos.
- Grande dependência do domínio:
  - Vantagens e desvantagens na identificação
- Exemplos:
  - Química/farmácia/medicina destacam-se pela sua excentricidade:
    - "3-4-benzoterminolanina" ou "Terminologite aguda"
  - Nas engenharias e ciências tecnológicas são mais variados mas "bem comportados":
    - "Sistema terminológico" ou "extractor de UT's"
  - Direito/geografia fazem parte do léxico corrente:
    - "Divórcio", "testemunha", "depressão", "bacia"

# Quanto à Sintaxe

- Normalmente consideram-se UT apenas sintagmas nominais
- Há várias estruturas sintácticas frequentes
  - N: rede
  - NA: rede neuronal
  - NpN: rede de telefone, ciclo de Krebs (EN!!)
  - NpNpNA: Sistema de reencaminhamento de Chamadas Automático
- Várias possibilidades de
  - Introduzir modificadores ou argumentos
  - Compor novas UT por variação:
    - Rede telefónica (apr. de sábado)

# Quanto à Sintaxe

- Há quem considere Verbos como UT válidas
  - mas essa não é a visão dominante.
- Normalmente as formas verbais resultam da transformação morfológica N→V da UT:

  Reencaminhamento de chamadas

  → reencaminhar (“as”|mod)* chamadas
- Por tudo isto, a detecção de variantes é uma área de investigação: apr. de sábado

# Recursos Terminológicos

- Onde armazenamos a Terminologia?
  - Tesauros: UMLS/MeSH
    - UT's + certas relações tradicionais
  - Ontologias (?)
    - UT's + certas relações específicas
  - Glossários
    - UT's e suas definições
  - Bases de Dados Terminológicas
    - Guardam as UT, equivalentes de Trad., Exemplos de Utilização, contextos / colocações, Inf. adminstrativa, etc.

# Recursos Terminológicos

- Apesar de ser uma representação lexical de um determinado conceito. uma UT não é guardada sozinha, mas sim:
  - Em contexto
  - Relacionada com outras UT
  - Relacionada com equivalentes noutros idiomas
  - Tal como os próprios conceitos
- Uma UT não tem valor isoladamente
  - Vale essencialmente pelas relações que estabelece

# Recursos Terminológicos

- Visão de Engenharia do Conhecimento
  - A UT como ponto de partida para a criação de um recurso de conhecimento...
  - ... e não só apenas chegada!
- A UT como “nó” de uma “rede”
  - poliedro irregular e instável de uma região semântica dinâmica e subjectiva
- Recursos gerados para determinados fins:
  - Um recurso terminológico não é uma representação final e última do conhecimento (se é que isso existe)
  - Podem/devem ser reformulados quando necessário

# Exemplos da utilização de UT

- Em Recolha de informação (D.Esp.?)
  - Indexação (“indexação controlada/conceptual”):
    - Se a UT é uma “representação lexical” de um conceito
    - Então pode ser uma representação mais fiel e compacta de um documento
  - Processamento de Expressões de Pesquisa
    - Tesauros para expansão / compactação
    - Substituição de uma UT por hiperónimo/co-hipónimo
  - Ordenação de resultados:
    - UT podem indicar com razoável precisão (?) o nível de especialização do documento: poucos exemplos
  - Auxílio à navegação e pesquisa de informação!
    - O utilizador pode “ver” os Termos “relevantes” e tentar reescrever a sua querie

# Exemplos da utilização de UT

- Trad. Assistida e Automática:
  - Para um tradutor humano uma das maiores dificuldades é a tradução de UT:
    - Reflectem o conhecimento do domínio que o tradutor (humano/automático) não tem!
    - Saber a Terminologia e as equivalências é fundamental (normalmente pergunta-se ao especialista)
  - Tradução automática as UT (Baseada Regras):
    - Importantes na Análise
    - Transferência não é normalmente ambígua (D.Esp)
    - Importantes na Geração

# Mais exemplos da utilização de UT

- Tarefas que envolvam:
  - análise rigorosa do texto
  - Necessidade de manter grande fidelidade semântica
    - Sumarização automática
    - ...
- Em quase todas as aplicações em que EM são importantes:
  - Shallow-parsing / Chunking/melhoramento “dinâmico” de parsers
  - QA
  - Reconhecimento de voz
    - Legendador automático
  - ...
- Ou seja, tarefas em que seja crítico assegurar que certas unidades lexicais (neste caso UT’s) não sejam segmentadas e sejam correctamente analisadas!

# Extracção de Terminologia

- Tarefa 1:
  - dado um texto num idioma:
    - Identificar as UT
    - Identificar possíveis variações
    - Identificar possíveis relações entre UT's
- Aquisição de Terminologia:
  - Se não há base terminológica prévia
  - Normalmente é desta que falamos
- Enriquecimento de Terminologia
  - Se se parte de uma base terminológica prévia

# Extracção de Terminologia

- Tarefa 2:
  - Tarefa 1 em textos de várias línguas
  - Obter terminologia alinhada
  - Normalmente textos comparáveis
  - Sub-problema de EBMT?
- Vamos focar:
  - Tarefa 1 - Aquisição de Terminologia

# Várias aproximações

- Há várias aproximações:
  - Gramaticais
    - Características morfológicas
    - sequências POS
  - Métodos estatísticos
    - Coesão lexical
    - Desvio relativo à norma
  - Híbridas
    - Combinações múltiplas
    - “micro-gramáticas”

# Várias aproximações

- Todas estas aproximações:
  - utilizam intuições e descrições linguísticas mais ou menos apuradas
  - têm requisitos de "pré-processamento" muito diferentes
  - Podem ou não prever retorno de um operador humano:
    - Há vários sistemas semi-automáticos que funcionam interactivamente

# Que tipo de evidência procurar?

- As UT são Unidades Lexicais internamente coesas e com forte influência contextual.
- Na sua detecção/exclusão podemos:
  - Procurar evidências internas:
    - morfologia / sintaxe
  - Procurar evidências externas:
    - Contextos descriminatórios / colocações
  - Combinar ambas as aproximações
    - “boundary rules”, exclusões internas
- Podem prever a possibilidade de variações morfo-sintácticas ou não…

# Aproximações Gramaticais

- Ideias base / Intuições:
  - As UT obedecem a uma gramática
    - Normalmente SN
    - EX: N (A|P N)+
      - Sistema de gestão de base de dados
      - Rede neuronal
      - Aparelho de golgi
  - Os contextos também possuem uma gramática
    - Mas não é tão normal pesquisar os contextos

# Aproximações Gramaticais

- Mais Ideias base / Intuições:
  - Identificar UT variantes sabendo o conjunto de regras que levam à sua formação:
  - Parte de um conjunto de UT base e aplica regras de composição -> gramática generativa
  - Pode utilizar / necessitar de retorno humano ou uma base de conhecimento inicial.
  - EX:
    - T (A|P)T)+
    - Regras de “analogia”:
      - TA[H1|M1] e TB[H1|M2] e TC[H2|M1] → TD[H2|M2]

# Aproximações Gramaticais

- Vantagens:
  - Boa precisão
  - Normalmente robustos
- Desvantagens:
  - complexidade de implementação
  - necessitam de bastante pré-processamento
  - Dificuldade de porte para outras línguas
  - Incapazes de detectar algo que não tenha sido previsto na gramática

# Aproximações Gramaticais

- Sistemas que exploram a Morfologia!
- Ideias base / Intuições:
  - Procurar sequências de palavras, que possuem uma determinada forma:
    - Sufixos: “proto*”
    - Prefixos: “*zoide”
    - Raizes: “sistema”
- Podem também usar alguma evidência externa simples:
  - Ex: Palavra anterior é artigo?
- mas não usam “análise sintáctica completa”

# Aproximações Gramaticais

- Vantagens:
  - Boa precisão em domínios “fechados”. Eg: Química
  - Dentro de um domínio, podem ser facilmente portados para línguas próximas (cognatos)
- Desvantagens:
  - Só são aplicáveis em certos domínios.
  - Necessitam de um investimento de estudo linguístico por cada domínio
  - Não são portáveis de domínio para domínio
  - Incapazes de detectar algo que não tenha sido previsto na “gramática”
- Podem ser uma boa maneira de “boot-strapping”

# Aproximações Estatísticas

- Sistemas que exploram informação estatística:
  - normalmente sobre o léxico mas pode haver lematização
- Ideias base / Intuições:
  1. As UT são unidades lexicais que pertencem a um determinado domínio, logo podem ser consideradas estatisticamente desviantes das unidades obtidas num "corpus padrão de linguagem comum"
     - Coeficiente de Dice sobre o N-Gramas extraídos
  2. As UT são unidades lexicais coesas: os seus constituintes co-ocorrem com frequências muito acima do "normal"
     - Informação Mútua
     - Coeficiente Z (Smadja)

# Aproximações Estatísticas

- Vantagens:
  - Portáveis entre línguas
  - Fáceis de programar
  - Rápidos (?)
- Desvantagens:
  - Muito ruidosos
    - Muita influência de fenómenos linguísticos recorrentes e de marcas de estilo
    - (2) pode obter todas as colocações, que não são UT!
  - Na prática, é difícil implementar:
    - Afinal, o que é um corpus padrão de linguagem geral?
- São uma boa aproximação quando não se tem alternativas que usam mais conhecimento explícito

# Aproximações Híbridas

- Ideias base / Intuições:
  - Os métodos anteriores parecem ser bons para lidar com certas características do problema – mas não com todas!
  - Tentar combinações que mantenham as vantagens de todos, e não as desvantagens
- Estes métodos são baseados normalmente em heurísticas que são verificadas num dado contexto aplicacional
- Podem levar a bons resultados mas normalmente:
  - não têm uma boa fundamentação teórica
  - podem requerer muita afinação manual
  - podem ser dependentes de um dado contexto/pressuposto não generalizável

# Aproximações Híbridas

- Pode ser praticamente tudo!
- Um caso concreto: o Corpógrafo
  - Alguma sintaxe
    - Regras de descrição (ou melhor de eliminação)
    - Análise de contextos próximos (palavra anterior)
  - Alguma “morfologia”:
    - Singularização para melhorar convergência
  - Alguma estatística
    - As UT seleccionadas ocorrem com uma frequência mínima

# O Corpógrafo – o contexto

- O contexto em que se realizam as tarefas de extracção é sempre importante!
- Corpógrafo: método orientado para pesquisas semi-automáticas em vários idiomas!
- É o primeiro passo para a construção de glossários/tesauros:
  - Permite a posterior extracção de definições
  - Permite a posterior identificação de relações
  - UT's funcionam como pontos de fixação do PLS
- NÃO foi ainda devidamente testado nem comparado!!

# O Corpógrafo

- A explicação do método irá ser feita mostrando o percurso de desenvolvimento realmente seguido
- Permitirá compreender melhor o seu funcionamento e assim:
  - Ser criticado/melhorado
  - Inspirar sistemas melhores/alternativos
- Por razões de reaproveitamento de dados os exemplos apresentados são em inglês
  - embora os resultados para português sejam equivalentes ou **superiores**.
- Em breve: Módulo Perl para utilização pública

# Uma aproximação muito simples

- Compilar N-gramas do corpus
- Perguntar ao utilizador se são UT
- Recolher os exemplos validados
- Vantagens:
  - Não são necessários recursos linguísticos
  - Não necessita de pré-processamento
  - Rápido e portável entre línguas
- Desvantagens
  - Demasiado ruidoso
  - Os utilizadores consideram inadequado
  - Não “aprende” nada com o retorno do utilizador

# Exemplo do corpus *Neurodemo*

- Domínio: neurologia
  - Textos tirados da web (pdf, word, html)
  - 6 idiomas (PT,EN,FR,ES,IT,DE)
  - Secção EM: 29192 àts.

| N-gram | # | F (%) |
|---|---|---|
| of the | 332 | 1.137 |
| in the | 243 | 0.832 |
| to the | 121 | 0.414 |
| the cell | 118 | 0.404 |
| from the | 71 | 0.243 |
| the brain | 65 | 0.222 |
| **nervous system** | 65 | 0.222 |
| on the | 59 | 0.202 |
| and the | 52 | 0.178 |
| of a | 51 | 0.174 |
| the neuron | 48 | 0.164 |
| the axon | 46 | 0.157 |
| **cell body** | 46 | 0.157 |
| is the | 42 | 0.143 |
| by the | 40 | 0.137 |
| in a | 40 | 0.137 |

**Muito ruidoso : apenas 2 UT em 15 n-grams!**

# Como melhorar os resultados?

- Vantagens: os resultados tão maus que podem ser facilmente melhorados.
- Podíamos criar regras sintácticas e morfológicas acerca das UT e seleccionar apenas N-gramas que as respeitem
- Contudo, como vimos, é muito difícil dizer o que é uma UT, e podemos ter de recorrer a algum pré-processamento: muito complicado...
- Mas é muito mais fácil dizer:
  - **o que é que NÃO É uma UT!**
  - Além de que é muito mais estável entre domínios

# Excluindo N-Gramas

- Definir 3 listas de átomos para exclusão de N-gramas
  - Lista dos não-inícios
    - àt. que não podem iniciar UT's
  - Lista dos não-fins
    - àt. que não podem terminar UT's
  - Lista dos não-incluídos
    - àt. que não podem estar incluídos em UT's
- Encontrar N-gramas que respeitem estas restrições
- Por razões de redundância, as UT devem:
  - aparecer no topo da lista dos N-gramas seleccionados.
- Aproximação similar a Merkel & Andersson, 2000

# E o que são estas listas?

- A maioria dos elementos destas listas são
  - preposições
  - pronomes
  - pontuação
  - certas palavras muito frequentes
- Facilmente compiladas por “tentativa-e-erro”
- Muito estáveis entre domínios
  - Hipótese até agora válida
  - Alteradas facilmente se necessário
- Fácil de adaptar a idiomas não algutinativos

# Restrições simples (top 20)

| N-gram (3110 found) | # | F (%) |
|---|---|---|
| nervous system | 65 | 0.222 |
| cell body | 46 | 0.157 |
| electrical activity | 39 | 0.133 |
| nerve cells | 37 | 0.126 |
| spinal cord | 34 | 0.116 |
| action potential | 32 | 0.109 |
| glial cells | 29 | 0.099 |
| synaptic cleft | 20 | 0.068 |
| plasma membrane | 16 | 0.054 |
| central nervous | 16 | 0.054 |
| action potentials | 14 | 0.047 |
| schwann cells | 14 | 0.047 |
| membrane proteins | 13 | 0.044 |
| nerve fibers | 13 | 0.044 |
| endoplasmic reticulum | 13 | 0.044 |
| nervous systems | 12 | 0.041 |
| developing circuits | 12 | 0.041 |
| amino acids | 12 | 0.041 |
| myelin sheath | 12 | 0.041 |
| nerve cell | 12 | 0.041 |

| N-gram (2208 found) | # | F (%) |
|---|---|---|
| central nervous system | 16 | 0.054 |
| peripheral nervous system | 8 | 0.027 |
| integral membrane proteins | 8 | 0.027 |
| nuclear pore complexes | 5 | 0.017 |
| name of glial | 5 | 0.017 |
| signaling between nerve | 5 | 0.017 |
| pattern of activity | 5 | 0.017 |
| nodes of ranvier | 4 | 0.013 |
| synthesis of proteins | 4 | 0.013 |
| induction of ltp/ltd | 4 | 0.013 |
| evoked nt secretion | 3 | 0.010 |
| can be divided | 3 | 0.010 |
| primary visual cortex | 3 | 0.010 |
| nmda receptor activation | 3 | 0.010 |
| – the messengers | 3 | 0.010 |
| action potential will | 3 | 0.010 |
| rate of transmission | 3 | 0.010 |
| primary cell walls | 3 | 0.010 |
| complexes of integral | 3 | 0.010 |
| refinement of neural | 3 | 0.010 |

# Restrições Simples

- Aumentamos a precisão (no topo)
  - Reduzimos o esforço de validação:
  - ~ 100 termos/hr
- Ainda temos alguns problemas:
  1. Algumas palavras frequentes/homógrafas ainda trazem muitos candidatos falsos (ex: “can”)
  2. Ocorrências divididas entre o Plural/Singular
  3. Termos encapsulados são ainda difíceis de separar
     - “nervous system” e “central **nervous system**”
  4. Não lidamos com nenhuma variação morfológica
  5. Não lidamos bem com UT incompletas

# Mais restrições...

- Objectivo: melhorar a precisão e resolver *alguns* dos problemas anteriores (1 & 2)
- Se “obrigarmos” a que os N-Gramas sejam SN, então a singularização é trivial em várias línguas
- Vamos impor 1 nova restrição: os N-Gramas têm de ser precedidos por certas palavras
  - Ex: “a”, “the”, “one”, “as”, etc.
  - É fácil impor restrições para “PT”, “ES”,”FR”, “IT”
  - É suficientemente simples de implementar

# Mais restrições: resultados

| N-gram (1304 found) | # | F (%) |
|---|---|---|
| cell body | 46 | 0.157 |
| nervous system | 35 | 0.119 |
| spinal cord | 31 | 0.106 |
| action potential | 30 | 0.102 |
| nerve cell | 26 | 0.089 |
| electrical activity | 22 | 0.075 |
| synaptic cleft | 20 | 0.068 |
| plasma membrane | 16 | 0.054 |
| glial cell | 16 | 0.054 |
| central nervous | 15 | 0.051 |
| myelin sheath | 12 | 0.041 |
| developing circuit | 11 | 0.037 |
| neural circuit | 10 | 0.034 |
| peripheral nervous | 9 | 0.030 |
| protein synthesis | 9 | 0.030 |
| endoplasmic reticulum | 8 | 0.027 |
| human brain | 8 | 0.027 |
| respiratory chain | 7 | 0.023 |
| schwann cell | 7 | 0.023 |
| nmda receptor | 7 | 0.023 |

| N-gram (943 found) | # | F (%) |
|---|---|---|
| central nervous system | 15 | 0.051 |
| peripheral nervous system | 9 | 0.030 |
| node of ranvier | 6 | 0.020 |
| integral membrane protein | 6 | 0.020 |
| synthesis of proteins | 4 | 0.013 |
| nuclear pore complex | 4 | 0.013 |
| induction of ltp/ltd | 4 | 0.013 |
| primary visual cortex | 3 | 0.010 |
| energy of atp | 3 | 0.010 |
| development of neural | 3 | 0.010 |
| rate of transmission | 3 | 0.010 |
| activation of nmda | 2 | 0.006 |
| evoked nt secretion | 2 | 0.006 |
| refinement of neural | 2 | 0.006 |
| xenopus retinotectal system | 2 | 0.006 |
| induction of ltp | 2 | 0.006 |
| activity-induced synaptic modification | 2 | 0.006 |
| cytochrome b gene | 2 | 0.006 |
| activity-dependent synaptic modification | 2 | 0.006 |
| developing neural circuit | 2 | 0.006 |

# Qual é a melhoria?

- Aumento da precisão
  - Torna-se mais fácil o processo de validação!
- Formas Plural/Singular convergiram
- Fácil de implementar
- “Multilingue”
- Muito rápido: como há exclusão de muitos N-Gramas, a ordenação é rápida!
  - 29K tokens em 2s - Intel P4

# Problemas por resolver

- Valor Abrangência: ainda desconhecido!
  - Precisamos de fazer testes!
- Problema do encapsulamento dos termos. Como resolver?
  - Simples: tentar primeiro achar N-Gramas maiores e partir só depois para o menores.
  - Custo: CPU...
- Problemas das variantes morfológicas:
  - Seria necessário implementar regras de transformação de equivalentes
  - Mais complicado... Mas não impossível

# O Método do Corpógrafo

- Algoritmo híbrido simples:
  - Fácil de implementar
  - Simples para utilizadores compreenderem
  - Execução em tempo linear - O(kN)
  - É possível alterar as restrições para adaptar a certos domínios
  - Pode ser portado para vários idiomas
  - Ainda melhorável, sem demasiado trabalho

# Extracção de Terminologia

- Conclusões:
  1. Terminologia é algo muito dinâmico
  2. Terminologia tem grande potencial aplicacional embora não tenha sido possível ainda explorar tudo com sucesso.
  3. A Extracção de Terminologia é por isso útil, mas é também uma tarefa dependente de um contexto
  4. Têm sido propostos vários métodos que possuem características mais adaptadas a certos contextos. Baseados em:
     - Gramáticas
     - Métodos estatísticos
     - Aproximações combinadas/ híbridas

# Extracção de Terminologia

- É possível implementar extractores de terminologia relativamente simples e com sucesso razoável (?)
- A extracção de terminologia é ainda uma área com alguma margem de desenvolvimento
- A terminologia ainda não é devidamente explorada em aplicações de PLN
  - Aplicações serão motivadoras de novos métodos que serão adaptados a um determinado contexto aplicacional (RI, TA, etc..)