Avaliação da extracção de relações semânticas entre palavras portuguesas a partir de um dicionário

Hugo Gonçalo Oliveira, Diana Santos, Paulo Gomes
hroliv@dei.uc.pt, diana.santos@sintef.no, pgomes@dei.uc.pt

STIL 2009

University of Coimbra
Faculty of Sciences and Technology
Department of Informatics Engineering

Knowledge and Intelligent Systems Laboratory
Cognitive and Media Systems Group
Centre of Informatics and Systems of the University of Coimbra

SINTEF
Natural Language Technologies Group

Ontologias lexicais

- Conjunto de termos e conceitos estruturados e ligados de acordo com relações léxico-semânticas
- Úteis para...
  - *Interpretar textos, determinar semelhanças entre conceitos, resposta automática a perguntas, tradução automática, geração de texto, pesquisa inteligente, estudos teóricos acerca da semântica de uma língua…*

Ontologias lexicais
Construção

- Manual
  - Menos propícia a erros
  - Difícil manutenção/actualização
- Semi-automática
  - A partir de dicionários
    - "autoridades" semânticas, vocabulário restrito
    - conhecimento geral
  - A partir de corpos
    - muita quantidade, rico em domínios específicos
    - texto muito variado e mais difícil de processar, não procuram abranger toda a língua

Ontologias lexicais
Exemplos

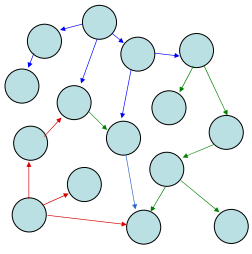
- Construção manual
  - Wordnet
  - Cyc
  - FrameNet
- Construção semi-automática
  - MindNet

Ontologias lexicais
Língua Portuguesa

- Construção manual
  - Tep (http://www.nilc.icmc.usp.br/tep2/)
  - WordNet.BR (http://www.nilc.icmc.usp.br/~arianidf/WordNet-BR.html)
  - WordNet.PT (http://cvc.instituto-camoes.pt/wordnet/)
  - MultiWordNet.PT (http://mwnpt.di.fc.ul.pt/)
- Construção semi-automática
  - PAPEL (http://www.linguateca.pt/PAPEL/)

PAPEL

- Palavras Associadas Porto Editora Linguateca



PAPEL

- Conjunto de (cerca de 200 mil) triplos relacionais entre termos portugueses

Grupo	Nomes	arg1, arg2	Num.	Exemplos
Sinonímia	SINONIMO_DE	qq,=arg1	80432	*(flexível, moldável)*
Hiperonímia	HIPERONIMO_DE	sub,sub	63455	*(planta, salva)*
Meronímia	PARTE_DE	sub,sub	14453	*(cauda, cometa)*
	PARTE_DE_ALGO_COM_PROPRIEDADE	sub,adj	3715	*(tampa, cobertor)*
	PROPRIEDADE_DE_ALGO_PARTE_DE	adj,sub	962	*(celular, célula)*
Causa	CAUSADOR_DE	sub,sub	1125	*(fricção, assadura)*
	CAUSADOR_DE_ALGO_COM_PROPRIEDADE	sub,adj	16	*(paixão, passional)*
	PROPRIEDADE_DE_ALGO_CAUSADOR_DE	adj, sub	515	*(reactivo, reacção)*
	ACCAO_QUE_CAUSA	v,sub	6424	*(limpar, purgação)*
	CAUSADOR_DA_ACCAO	sub,v	39	*(gases, fumigar)*
Produtor	PRODUTOR_DE	sub,sub	932	*(romãzeira, romã)*
	PRODUTOR_DE_ALGO_COM_PROPRIEDADE	sub,adj	31	*(sublimação, sublimado)*
	PROPRIEDADE_DE_ALGO_PRODUTOR_DE	adj,sub	348	*(fotógeno, luz)*
Fim	FINALIDADE_DE	sub,sub	2095	*(passagem,a,catedrático, agregação)*
	FINALIDADE_DE_ALGO_COM_PROPRIEDADE	sub,adj	23	*(enumeração, enumerativo)*
	ACCAO_FINALIDADE_DE	v,sub	5640	*(fitzer_rir, comédia)*
	ACCAO_FINALIDADE_DE_ALGO_COM_PROPRIEDADE	v,adj	255	*(corrigir, correccional)*
	MANEIRA_POR_MEIO_DE	adv,sub	1433	*(timidamente, timidez)*
Lugar	LOCAL_ORIGEM_DE	sub,sub	768	*(Japão, japonês)*
Propriedade	PROPRIEDADE_DE_ALGO_REFERENTE_A	adj,sub	3700	*(dinâmico, movimento)*
	PROPRIEDADE_DO_QUE	adj,v	17028	*(diplomado, possuir_diploma)*

PAPEL

- Extraídos de forma semi-automática, com base em gramáticas semânticas (específicas para cada relação) que incluem padrões indicadores

Padrão	Relação associada
`tipo|género|classe|forma de`	Hiperonímia
`parte|membro de`	Meronímia
`que causa|provoca|origina`	Causa
`usado|utilizado para`	Objectivo
uma palavra ou lista de palavras	Sinonímia

PAPEL

- Relações genéricas (REL_GEN) e também mais específicas (REL_ESP), com base nas categorias gramaticais dos argumentos, pré-definidas da seguinte forma:

```
REL_GEN {
cat1:cat2 * REL_ESP_1:REL_ESP_1_inv;
cat1:cat3 * REL_ESP_2:REL_ESP_2_inv;
…
}
```

```
PARTE {
nome:nome * PARTE_DE:INCLUI;
nome:adj * PARTE_DE_ALGO_COM_PROPRIEDADE:PROPRIEDADE_DE_ALGO_QUE_INCLUI;
adj:nome * PROPRIEDADE_DE_ALGO_PARTE_DE:INCLUI_ALGO_COM_PROPRIEDADE;
}

CAUSA{
nome:nome * CAUSADOR_DE:RESULTADO_DE;
nome:verbo * CAUSADOR_DA_ACCAO:ACCAO_RESULTADO_DE;
nome:adj * CAUSADOR_DE_ALGO_COM_PROPRIEDADE:PROPRIEDADE_DE_ALGO_RESULTADO_DE;
adj:nome * PROPRIEDADE_DE_ALGO_QUE_CAUSA:RESULTADO_DE_ALGO_COM_PROPRIEDADE;
verbo:nome * ACCAO_QUE_CAUSA:RESULTADO_DA_ACCAO;
}
```


PAPEL

- Extracção

1. **cometa, s. m.**
   astro geralmente constituído por núcleo, cabeleira e cauda

2.
```
[RAIZ]
  [QUALQUERCOISA]
    > [astro]
    [QUALQUERCOISA]
      > [geralmente]
  [PADRAO_CONSTITUIDO]
    [VERBO_PARTE_PP]
      > [constituído]
    [PREP]
      > [por]
  [ENUM_PARTE]
    [PARTE_DE]
      > [núcleo]
    [VIRG]
      > [,]
    [ENUM_PARTE]
      [PARTE_DE]
        > [cabeleira]
      [CONJ]
        > [e]
      [PARTE_DE]
        > [cauda]
```

3. núcleo PARTE_DE cometa
   cabeleira PARTE_DE cometa
   cauda PARTE_DE cometa

PAPEL
Mais exemplos de extracção

cólera, s. f. grave doença epidémica, contagiosa, que provoca diarreia, vómitos e cólicas, causada por um bacilo (vibrião) e também designada por cólera-morbo e cólera-asiática, mordexim ou mordixim
- --> diarreia RESULTADO_DE cólera
- --> vómitos RESULTADO_DE cólera
- --> cólicas RESULTADO_DE cólera

clarabóia, s. f. janela ou fresta por onde entra a luz num aposento
- --> fresta HIPERONIMO_DE clarabóia
- --> janela HIPERONIMO_DE clarabóia

tapadura, s. f. objecto próprio para tapar ou cobrir
- --> tapar ACCAO_FINALIDADE_DE tapadura
- --> cobrir ACCAO_FINALIDADE_DE tapadura

vínculo, s. m. tudo o que serve para prender ou atar
- --> prender ACCAO_FINALIDADE_DE vínculo
- --> atar ACCAO_FINALIDADE_DE vínculo

PAPEL

- Ajuste de relações:
  - Tudo para tipo directo
    - *manga* INCLUI *punho* >> *punho* PARTE_DE *manga*
    - *dor* RESULTADO_DE *distensão* >> *distensão* CAUSADOR_DE *dor*
  - Lematização dos argumentos
  - Correcção do nome das relações
    - *loucura* ACCAO_QUE_CAUSA *desvario* >> *loucura* CAUSADOR_DE *desvario*

Avaliação
Sinonímia

- Utilização do TeP como recurso dourado
- Remoção das relações com termos que não são comuns a ambos os recursos
  - 50% do PAPEL no TeP, 39% do TeP no PAPEL
- Expansão: (A SINONIMO_DE B) e (B SINONIMO_DE C) >> (A SINONIMO_DE C)
  - 19% do PAPEL no TeP, 90% do TeP no PAPEL
  - Incoerências

    A=*ruína*, B=*queda*, C=*habilidade*

    >> *ruína* SINONIMO_DE *habilidade*

Avaliação
Restantes relações

- Avaliação das restantes relações
  1. Tradução das relações para "linguagem natural"
  2. Busca no CETEMPúblico, através do AC/DC

Relação	Certa?	Justificação
língua HIPERONIMO_DE *italiano*	Sim	*As **línguas latinas, como o italiano** ou o português, tornam-se mais fáceis por causa das vogais.*
arbusto PARTE_DE *floresta*	Sim	*A **floresta é um conjunto de árvores, arbustos** e ervas de várias qualidades e tamanhos.*
cólera CAUSADOR_DE *diarreia*	Sim	*A **cólera provoca fortes diarreias** e vómitos e pode levar à desidratação e, consequentemente, à morte em poucas horas.*
oliveira PRODUTOR_DE *azeitona*	Sim	*Também a quantidade e tamanho das **azeitonas produzidas por uma oliveira** biológica é inferior, já que não são utilizados compostos de azoto que ajudam a planta a crescer.*
recrutamento FINALIDADE_DE *inspecção*	Sim	*Menos de metade dos jovens entre os 20 e os 22 anos apresentaram-se às **inspecções para recrutamento**, revelou o ministro da Defesa.*
músico PARTE_DE *música*	Não	*... um espectáculo baseado na obra "Cantos de Maldoror", de Lautréamont, com **música composta pelo músico** inglês Steven Severin...*
fim FINALIDADE_DE *sempre*	Não	*Sicília aponta **sempre para o fim** do dia, para o fim da luz.*

Avaliação
Usando o AC/DC

- Remoção das relações com argumentos inexistentes no CETEMPúblico, a partir das listas de frequência dos lemas
- Criação de padrões simples de interrogação:

```
PRED_CAUSA_A ::=
  [lema="poder|ir|ter|dever"]*
  [lema="causar|provocar|originar|suscitar|m
  otivar|resultar|produzir|gerar"]
  [pos="ADV.*"]*
INSTANCIA_DE ::=
  [lema="tipo|forma|instância|género"]
  [lema="de"]
```


Resultados da primeira avaliação
Apenas com padrões simples no CETEMPúblico v. 4.0

- Primeiros resultados

Relação	Relações c/ args no CETEMPúblico	%	Amostra	%	Encontradas	%
Hiperonímia	40,079	63%	3,145	8%	560	18%
Meronímia	3,746	35%	2,343	63%	521	22%
Causa	557	50%	557	100%	20	4%
Produtor	414	44%	414	100%	12	3%
Finalidade	1,718	59%	1,718	100%	173	10%

- Algumas relações correctas, mas não encontradas:
  - fruto HIPERONIMO_DE alperce
  - algoritmia PARTE_DE matemática
  - ausência CAUSADOR_DE saudade
  - aquecimento FINALIDADE_DE salamandra
  - …

Resultados da primeira avaliação
Apenas com padrões simples no CETEMPúblico v. 4.0 (cont.)

- Muitas das relações extraídas, por vezes até do foro etimológico, não aparecem naturalmente em texto noticioso
  - `China LOCAL_ORIGEM_DE chinês`
  - `forma_de_célula PARTE_DE_ALGO_COM_PROPRIEDADE celuliforme`
- Os padrões não cobrem por enquanto muitos casos
  - `viga PARTE_DE ponte`
  - *par=ext423909-soc-92a-1*: Ferreira do Amaral conta poder adjudicar o reforço da **ponte** (com tirantes e uma nova **viga**) e a construção por privados da nova ferrovia dentro de ano e meio
- Vários falsos positivos, sobretudo nos casos de padrões muito genéricos

Comentários finais

- Ponto de partida para projectos futuros
  - Novas avaliações
  - Adição de informação de fiabilidade às relações
  - Estudo quantitativo da polissemia
  - Comparação com outros recursos
  - Integração com outros recursos e outras metodologias
  - Seu uso para interrogação de corpos (AC/DC)
- O PAPEL é completamente livre: todos podem fazer o que quiserem com ele, melhorá-lo, e redistribuí-lo
  - Há muito para melhorar
  - Contamos com novas versões e avaliações em colaboração

Agradecimentos

- O PAPEL foi desenvolvido no âmbito da Linguateca, sob o contrato POSC/339/1.3/C/NAC, co-financiado pelo governo português, pela União Europeia (FEDER e FSE), pela UMIC e pela FCCN.

- Hugo Gonçalo Oliveira é financiado pela FCT, bolsa SFRH/BD/44955/2008.
- Agradecemos ainda ao grupo de I&D da Porto Editora o apoio na construção do PAPEL.

Muito em breve...

- Diana Santos, Anabela Barreiro, Luís Costa, Cláudia Freitas, Paulo Gomes, Hugo Gonçalo Oliveira, José Carlos Medeiros & Rosário Silva. "O papel das relações semânticas em português: Comparando o TeP, o MWN.PT e o o PAPEL". In *XXV Encontro Nacional da Associação Portuguesa de Linguística* (Lisboa, Portugal, 22-24 de Outubro de 2009).