Extracção e organização de relações semânticas

Hugo Gonçalo Oliveira
hroliv@dei.uc.pt

Escola de Verão
CLUP, 1 de Julho de 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

Introdução

- Processamento de Linguagem Natural
  - Fonologia
  - Morfologia
  - Sintaxe
  - **Semântica**
  - Pragmática
  - Discurso

Introdução

- Semântica lexical: palavras e os seus significados
- Relações léxico-semânticas: relações entre termos ou conceitos.
  - Sinonímia: termos diferentes têm o mesmo significado (*casa, domicílio, habitação*)
  - Hiponímia: um conceito é uma subclasse de outro (*carro é um veículo*)
  - Meronímia: um conceito é parte de outro (*roda de um carro*)
  - Causa: um conceito origina outro (*gripe causa febre*)
  - …

Recursos lexicais

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

Recursos lexicais
Construção manual

- Wordnet: baseado em princípios psicolinguísticos
- Conceitos representados por listas de sinónimos (*synsets*)
- Relações entre conceitos

- Distribuído gratuitamente
- Acessível a partir de http://wordnet.princeton.edu/

Recursos lexicais
Construção manual

- *bird* (Princeton WordNet 3.0)

Noun

- S: (n) bird (warm-blooded egg-laying vertebrates characterized by feathers and forelimbs modified as wings)
  - *direct hyponym* / *full hyponym*
    - S: (n) dickeybird, dickey-bird, dickybird, dicky-bird (small bird; adults talking to children sometimes use these words to refer to small birds)
    - S: (n) cock (adult male bird)
    - S: (n) hen (adult female bird)
    - S: (n) nester (a bird that has built (or is building) a nest)
    - S: (n) night bird (any bird associated with night: owl; nightingale; nighthawk; etc)
    - S: (n) bird of passage (any bird that migrates seasonally)
    - S: (n) passerine, passeriform bird (perching birds mostly small and living near the ground with feet having 4 toes arranged to allow for gripping the perch; most are songbirds; hatchlings are helpless)
    - S: (n) bird of prey, raptor, raptorial bird (any of numerous carnivorous birds that hunt and kill other animals)
    - S: (n) gallinaceous bird, gallinacean (heavy-bodied largely ground-feeding domestic or game birds)
    - S: (n) parrot (usually brightly colored zygodactyl tropical birds with short hooked beaks and the ability to mimic sounds)
    - S: (n) trogon (forest bird of warm regions of the New World having brilliant lustrous plumage and long tails)
    - S: (n) aquatic bird (wading and swimming and diving birds of either fresh or salt water)
    - S: (n) twitterer (a bird that twitters)
    - …
  - *part meronym*
  - *member holonym*
  - *direct hypernym* / *inherited hypernym* / *sister term*
  - *domain term category*
  - *derivationally related form*
- S: (n) bird, fowl (the flesh of a bird or fowl (wild or domestic) used as food)
- S: (n) dame, doll, wench, skirt, chick, bird (informal terms for a (young) woman)
- S: (n) boo, hoot, Bronx cheer, hiss, raspberry, razzing, razz, snort, bird (a cry or noise made to express displeasure or contempt)
- S: (n) shuttlecock, bird, birdie, shuttle (badminton equipment consisting of a ball of cork or rubber with a crown of feathers)

Verb

- S: (v) bird, birdwatch (watch and study birds in their natural habitat)

Recursos lexicais
Construção automática

- A construção manual é menos propícia a erros, mas é muito trabalhosa e de difícil manutenção e actualização
- Extracção automática de informação semântica em…
- Dicionários
  - Prós: “autoridades” semânticas, vocabulário restrito
  - Contras: conhecimento demasiado geral
- *Corpora*
  - Prós: muita quantidade, rico em domínios específicos
  - Contras: texto muito variado e mais difícil de processar, não abrangem toda a linguagem

Recursos lexicais
Construção automática

- Detecção de padrões léxico-sintácticos (Chodorow et al. 1985, Hearst 1992)
  - Simples de implementar
  - Bons resultados na extracção de hiperonímia
- Gramáticas semânticas específicas para os recursos a processar (Alshawi 1989)
- *Parsers* genéricos (Montemagni & Vanderwende 1992)
  - Melhor na identificação de características distintivas de co-hipónimos
  - Adaptável a qualquer tipo de texto

Recursos lexicais
Construção automática

- MindNet: metodologia para adquirir, estruturar, aceder e explorar informação semântica em recursos estruturados (e não só).
- Caminhos de relações para inferir semelhança
  - Utilização de um tesauro para identificar caminhos comuns entre sinónimos e hiperónimos
- Proprietário (Microsoft)
- Acessível a partir de http://stratus.research.microsoft.com/mnex/Main.aspx

Recursos lexicais

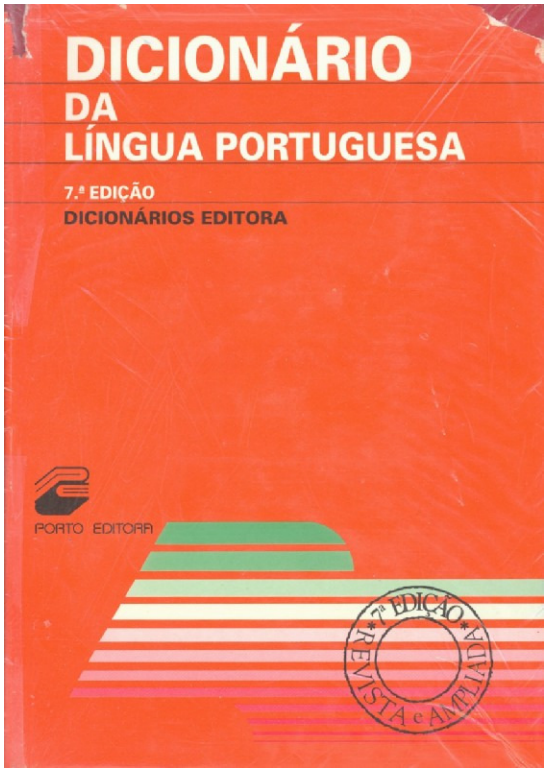
- MindNet: Caminhos de *bird* para *parrot*
  - bird ← Hyp ← parrot
  - bird → Mod → parrot
  - bird → Equiv → parrot
  - bird ← Tsub ← include → Tobj → parrot
  - bird → Attrib → flightless ← Attrib ← parrot
  - bird ← Tsub ← deplete → Tsub → parrot
  - bird → PrepRel(as) → kea → Hyp → parrot
  - bird ← Hyp ← macaw → Equiv → parrot
  - bird → PrepRel(as) → species → PrepRel(of) → parrot
  - bird → Attrib → flightless ← Attrib ← kakapo → Hyp → parrot

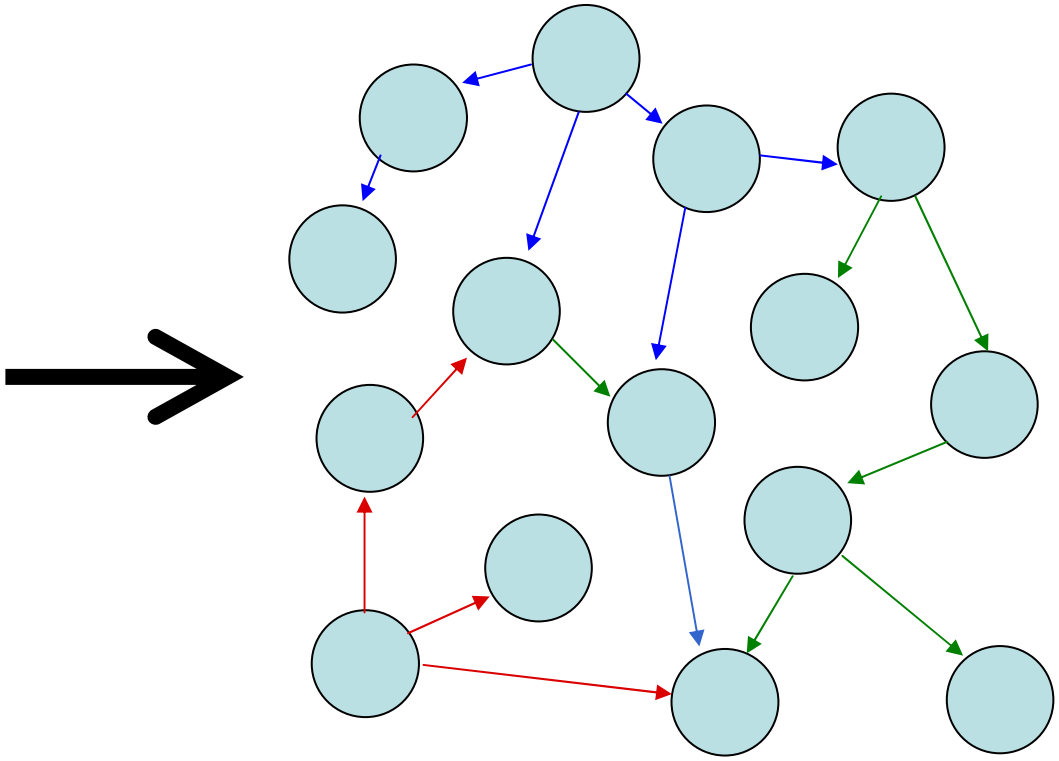
Recursos lexicais
Para a Língua Portuguesa

- Tep (http://www.nilc.icmc.usp.br/tep2/)
  - papagaio (Substantivo)
    - 1. papagaio, pipa, quadrado
    - 2. papagaio, compadre, patinho
    - 3. papagaio, falador, grulha, palrador, prosista, tagarela, tarelo, vozeiro
- 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/)
- PAPEL (http://www.linguateca.pt/PAPEL)

PAPEL

- Palavras Associadas Porto Editora Linguateca
- 1ª versão disponível a 17 de Agosto de 2009





PAPEL

- Conjunto de (cerca de 200 mil) relações entre termos

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*, *coberto*)
	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	(*fazer_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 (e subespecificações com base nas categorias gramaticais dos argumentos) pré-definidas

```
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

- Ajuste de relações:
  - Passagem para o 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 destes recursos

- Não é muito comum, principalmente quando a construção foi manual
- Inspiração em métodos para avaliação de ontologias de domínio, mas…
- As ontologias lexicais têm características diferentes!
- Ainda assim…

Avaliação destes recursos

- Utilização do Wordnet como recurso dourado (Hearst 1992, Nichols et al. 2005)
- Avaliações independentes ao WordNet
  - validação automática dos synsets, com recurso a um dicionário (Raman and Bhattacharyya 2008)
- Para o MindNet,
  - manualmente revistas 250 relações, escolhidas de forma a poder generalizar (Richardson et al. 1993)
  - Avaliação do mecanismo de inferência (Richardson 1997)

Presente
Avaliação do PAPEL

- Avaliação de 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

Presente
Avaliação do PAPEL

- Avaliação das restantes relações
  - 1. Tradução das relações para língua natural
  - 2. Busca no CETEMPúblico

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	*Sicilia aponta* ***sempre para o fim*** *do dia, para o fim da luz.*

Presente
Avaliação do PAPEL

- Resultados

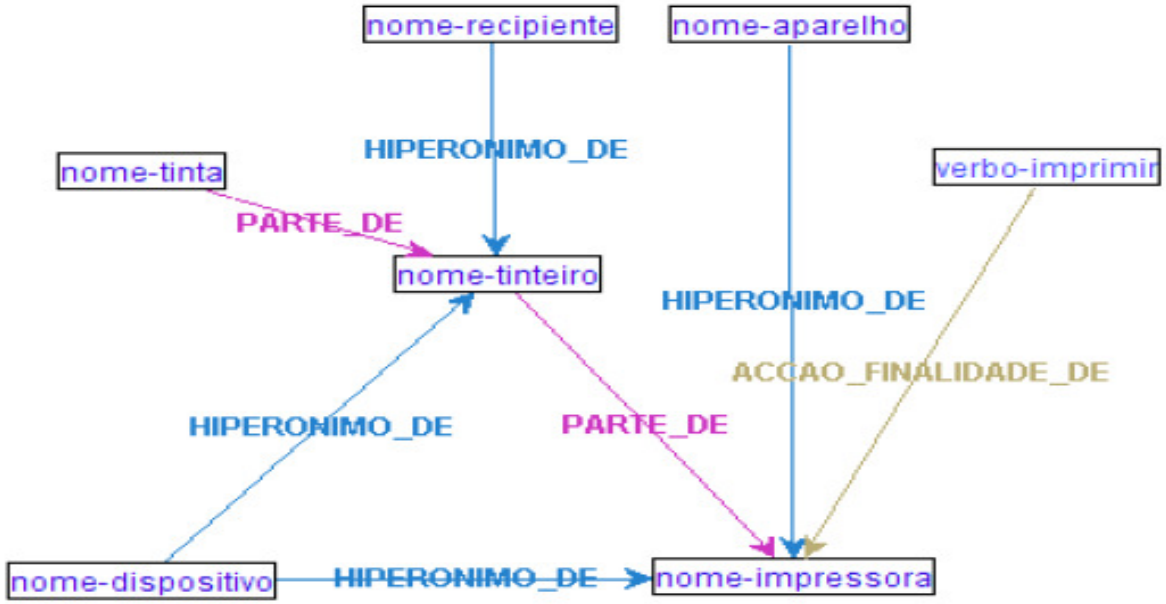
Relação	Relações c/ args no CETEMPúblico	%	Amostra	%	Encontradas	%
Hiperonímia	40,079	63%	3,145	8%	560	18%
Meronímia	3,746	35%	2,343	63%	521	22%
Causa	557	50%	557	100%	20	4%
Produtor	414	44%	414	100%	12	3%
Finalidade	1,718	59%	1,718	100%	173	10%

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

Presente

Navegação/visualização do PAPEL

dispositivo HIPERONIMO_DE impressora
aparelho HIPERONIMO_DE impressora
imprimir ACCAO_FINALIDADE_DE impressora
tinteiro PARTE_DE impressora
dispositivo HIPERONIMO_DE tinteiro
recipiente HIPERONIMO_DE tinteiro
tinta PARTE_DE tinteiro



Presente
Utilização do PAPEL

- Geração automática de poesia

```
se eu fosse dispositivo era joystick
uma flor terá olho-de-gato
se eu fosse agenda era calepino
detonação de cargas é resultado de detonador

como olho-de-gato sem uma flor
como um laser produziria um raio
furta-fogo perderia dispositivo
se eu for umas impressoras serás tinteiro

como olho-de-gato sem a flor
se eu fosse membro da flor era olho-de-gato
se ela fosse detonador provocasse detonação de cargas

um pulmão procura dispositivo
xipo é uma espécie do cinto
como injector sem aparelho de sulfatar
```

Futuro

- Melhorias ao PAPEL
- Adaptação da metodologia a outros dicionários (e.g. Wikcionário)
- Junção dos resultados numa única ontologia lexical: Onto.PT
- Desenvolvimento de ferramentas para extrair informação semântica a partir de *corpora* e enriquecer o recurso em domínios específicos
- Novas avaliações
- Novidades em http://eden.dei.uc.pt/~hroliv/phd.html

Extracção e organização de relações semânticas

Hugo Gonçalo Oliveira
hroliv@dei.uc.pt

Escola de Verão
CLUP, 1 de Julho de 2009

Obrigado!