

Extracção automática de conhecimento léxico-semântico a partir da Wikipédia

Hugo Gonçalo Oliveira¹, Hernani Costa, Paulo Gomes

hroliv@dei.uc.pt, hpcosta@student.dei.uc.pt, pgomes@dei.uc.pt

Cognitive & Media Systems Group
CISUC, Universidade de Coimbra

Braga, 9 Setembro 2010

¹ financiado pela bolsa da FCT SFRH/BD/44955/2008



- 1 Introdução
 - Ontologias lexicais
 - Enquadramento
- 2 Arquitectura do sistema
 - Módulo de extracção
- 3 Experimentação
 - Resultados
 - Avaliação
- 4 Notas finais



Introdução

- Hoje em dia...
 - ▶ Há uma grande quantidade de informação escrita



Introdução

- Hoje em dia...
 - ▶ Há uma grande quantidade de informação escrita
- Para tirar partido desta informação, as aplicações...
 - ▶ Necessitam de interpretar a linguagem natural



Introdução

- Hoje em dia...
 - ▶ Há uma grande quantidade de informação escrita
- Para tirar partido desta informação, as aplicações...
 - ▶ Necessitam de interpretar a linguagem natural
 - ▶ Requerem um melhor acesso a informação léxico-semântica



Introdução

- Hoje em dia...
 - ▶ Há uma grande quantidade de informação escrita
- Para tirar partido desta informação, as aplicações...
 - ▶ Necessitam de interpretar a linguagem natural
 - ▶ Requerem um melhor acesso a informação léxico-semântica
 - ▶ Estruturada em bases de conhecimento lexical (pe. ontologias lexicais)



Ontologias lexicais

- Tal como:
 - ▶ Princeton WordNet [Fellbaum, 1998]



Ontologias lexicais

- Tal como:
 - ▶ Princeton WordNet [Fellbaum, 1998]
- Definição:
 - ▶ Ontologia + léxico [Hirst, 2004]



Ontologias lexicais

- Tal como:
 - ▶ Princeton WordNet [Fellbaum, 1998]
- Definição:
 - ▶ Ontologia + léxico [Hirst, 2004]
 - ▶ Estruturadas em palavras e no seu significado



Ontologias lexicais

- Tal como:
 - ▶ Princeton WordNet [Fellbaum, 1998]
- Definição:
 - ▶ Ontologia + léxico [Hirst, 2004]
 - ▶ Estruturadas em palavras e no seu significado
 - ▶ Abranger toda uma língua



Ontologias lexicais

- Tal como:
 - ▶ Princeton WordNet [Fellbaum, 1998]
- Definição:
 - ▶ Ontologia + léxico [Hirst, 2004]
 - ▶ Estruturadas em palavras e no seu significado
 - ▶ Abranger toda uma língua
 - ▶ Sem domínio específico



Ontologias lexicais

- Tal como:
 - ▶ Princeton WordNet [Fellbaum, 1998]
- Definição:
 - ▶ Ontologia + léxico [Hirst, 2004]
 - ▶ Estruturadas em palavras e no seu significado
 - ▶ Abranger toda uma língua
 - ▶ Sem domínio específico
- Muitas vezes criadas de forma manual...
 - ▶ Construção e manutenção muito trabalhosas!



Construção automática

- Dicionários
 - ▶ Estruturados em palavras e significados
 - ▶ Vocabulário simples, quase previsível



Construção automática

- Dicionários
 - ▶ Estruturados em palavras e significados
 - ▶ Vocabulário simples, quase previsível
 - ▶ Limitados e estáticos
 - ▶ Nem sempre disponíveis



Construção automática

- Dicionários
 - ▶ Estruturados em palavras e significados
 - ▶ Vocabulário simples, quase previsível
 - ▶ Limitados e estáticos
 - ▶ Nem sempre disponíveis
- Corpos
 - ▶ Grande disponibilidade
 - ▶ Grande quantidade/variedade de conhecimento



Construção automática

- Dicionários
 - ▶ Estruturados em palavras e significados
 - ▶ Vocabulário simples, quase previsível
 - ▶ Limitados e estáticos
 - ▶ Nem sempre disponíveis
- Corpos
 - ▶ Grande disponibilidade
 - ▶ Grande quantidade/variedade de conhecimento
 - ▶ Poucas restrições (síntaxe e vocabulário)



Construção automática

- Dicionários
 - ▶ Estruturados em palavras e significados
 - ▶ Vocabulário simples, quase previsível
 - ▶ Limitados e estáticos
 - ▶ Nem sempre disponíveis
- Corpos
 - ▶ Grande disponibilidade
 - ▶ Grande quantidade/variedade de conhecimento
 - ▶ Poucas restrições (síntaxe e vocabulário)
- Enciclopédias
 - ▶ Conhecimento sobre o mundo e saber humano
 - ▶ Descrições mais extensas



Construção automática

- Dicionários
 - ▶ Estruturados em palavras e significados
 - ▶ Vocabulário simples, quase previsível
 - ▶ Limitados e estáticos
 - ▶ Nem sempre disponíveis
- Corpos
 - ▶ Grande disponibilidade
 - ▶ Grande quantidade/variedade de conhecimento
 - ▶ Poucas restrições (sintaxe e vocabulário)
- Enciclopédias
 - ▶ Conhecimento sobre o mundo e saber humano
 - ▶ Descrições mais extensas
 - ▶ **Wikipédia**: criada colaborativamente, disponível gratuitamente...



Enquadramento

- Objectivo
 - ▶ Extracção automática de relações semânticas a partir da Wikipédia portuguesa



Enquadramento

- Objectivo
 - ▶ Extracção automática de relações semânticas a partir da Wikipédia portuguesa
- Abordagem
 - ▶ Parser + Gramáticas semânticas
 - ▶ *Abstracts* da Wikipédia



Enquadramento

- Objectivo
 - ▶ Extracção automática de relações semânticas a partir da Wikipédia portuguesa
- Abordagem
 - ▶ Parser + Gramáticas semânticas
 - ▶ *Abstracts* da Wikipédia
- Integração futura no Onto.PT [Gonçalo Oliveira and Gomes, 2010]
 - ▶ Construção automática de uma ontologia lexical para o português



Sistema de extracção

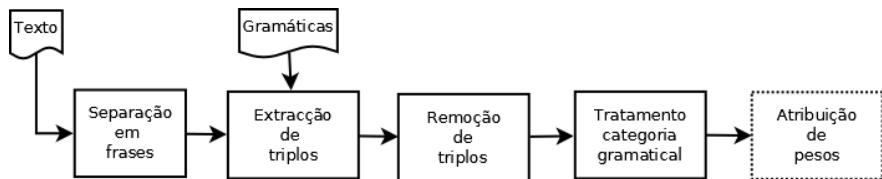


Figure: Os módulos do sistema de extracção.

²<http://code.google.com/p/pen/>

³<http://opennlp.sourceforge.net>

Sistema de extracção

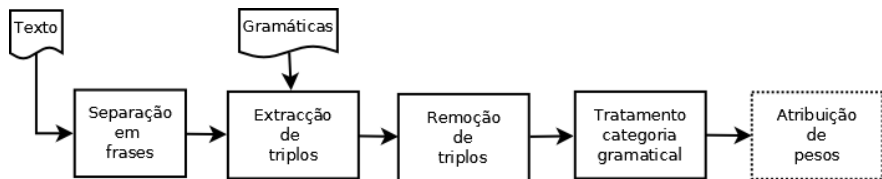


Figure: Os módulos do sistema de extracção.

- PEN² para analisar o texto de acordo com as gramáticas

²<http://code.google.com/p/pen/>

³<http://opennlp.sourceforge.net>

Sistema de extracção

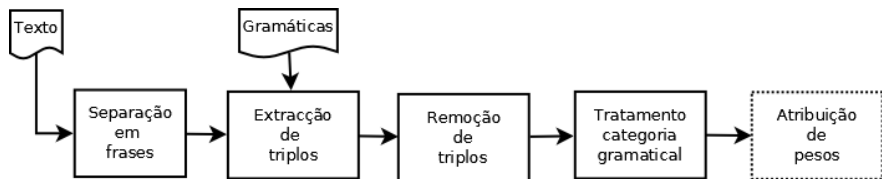


Figure: Os módulos do sistema de extracção.

- PEN² para analisar o texto de acordo com as gramáticas
- OpenNLP³, treinado para português para indicar categorias gramaticais e lematização.

²<http://code.google.com/p/pen/>

³<http://opennlp.sourceforge.net>

Extracção de triplos

- Extracção com base em padrões:



Extracção de triplos

- Extracção com base em padrões:
 - ▶ O automobilismo (também conhecido como corridas de automóveis ou desporto motorizado) é um desporto...
 - *automobilismo* SINONIMO_DE *corridas_de_automóveis*
 - *automobilismo* SINONIMO_DE *desporto_motorizado*
 - *desporto* HIPERONIMO_DE *automobilismo*



Extracção de triplos

- Extracção com base em padrões:
 - ▶ O automobilismo (também conhecido como corridas de automóveis ou desporto motorizado) é um desporto...
 - *automobilismo* SINONIMO_DE *corridas_de_automóveis*
 - *automobilismo* SINONIMO_DE *desporto_motorizado*
 - *desporto* HIPERONIMO_DE *automobilismo*
- Extracção com base em termos compostos ([N ADJ], [N de|do|da|com|para N]):



Extracção de triplos

- Extracção com base em padrões:
 - ▶ O automobilismo (também conhecido como corridas de automóveis ou desporto motorizado) é um desporto...
 - *automobilismo* SINONIMO_DE *corridas_de_automóveis*
 - *automobilismo* SINONIMO_DE *desporto_motorizado*
 - *desporto* HIPERONIMO_DE *automobilismo*
- Extracção com base em termos compostos ([N ADJ], [N de|do|da|com|para N]):
 - ▶ O miocárdio é um músculo cardíaco, isto é, à própria parede do coração
 - *músculo* HIPERONIMO_DE *músculo_cardíaco*
 - *parede* HIPERONIMO_DE *parede_do_coração*



Preparação

- Abstracts da Wikipédia Portuguesa (37,898 frases)



Preparação

- Abstracts da Wikipédia Portuguesa (37,898 frases)
 - ▶ Remoção de entradas de categorias menos centrais na construção de uma ontologia lexical



Preparação

- Abstracts da Wikipédia Portuguesa (37,898 frases)
 - ▶ Remoção de entradas de categorias menos centrais na construção de uma ontologia lexical
 - ▶ Apenas entradas das categorias (taxonomia da DBPedia [Bizer et al., 2009])
 - ★ *Species, AnatomicalStructure, ChemicalCompound, Disease, Currency, Drug, Activity, Language, MusicGenre, Colour, EthnicGroup e Protein*



Preparação

- Abstracts da Wikipédia Portuguesa (37,898 frases)
 - ▶ Remoção de entradas de categorias menos centrais na construção de uma ontologia lexical
 - ▶ Apenas entradas das categorias (taxonomia da DBPedia [Bizer et al., 2009])
 - ★ *Species, AnatomicalStructure, ChemicalCompound, Disease, Currency, Drug, Activity, Language, MusicGenre, Colour, EthnicGroup e Protein*
- Gramáticas semânticas
 - ▶ Com padrões que frequentemente indicam as relações:
 - ★ Sinonímia, Hiperonímia, Parte, Causa, Finalidade



Resultados

Relação	Extraídos		Exemplos
	Total	S/rep	
Hiperonímia TC	24.367	16.228	(<i>desordem, desordem_cerebral</i>) (<i>átomo, átomo_de_carbono</i>)
Hiperonímia	31.254	29.563	(<i>desporto, automobilismo</i>) (<i>estilo_de_música, folk</i>)
Sinonímia	11.872	11.862	(<i>inglês_antigo, anglo-saxão</i>) (<i>estupro, violação</i>)
Parte	1.321	1.287	(<i>jejuno, instestino</i>) (<i>rolas, columbidae</i>)
Finalidade	777	743	(<i>amoxicilina, tratamento_de_infeções</i>) (<i>construção, terracota</i>)
Causador	559	520	(<i>parasita, doença</i>) (<i>doença_neuromuscular, fadiga</i>)

Resultados

Relação	Extraídos		Exemplos
	Total	S/rep	
Hiperonímia TC	24.367	16.228	(<i>desordem,desordem_cerebral</i>) (<i>átomo,átomo_de_carbono</i>)
Hiperonímia	31.254	29.563	(<i>desporto,automobilismo</i>) (<i>estilo_de_música,folk</i>)
Sinonímia	11.872	11.862	(<i>inglês_antigo,anglo-saxão</i>) (<i>estupro,violação</i>)
Parte	1.321	1.287	(<i>jejuno,instestino</i>) (<i>rolas,columbidae</i>)
Finalidade	777	743	(<i>amoxicilina,tratamento_de_infeções</i>) (<i>construção, terracota</i>)
Causador	559	520	(<i>parasita, doença</i>) (<i>doença_neuromuscular,fadiga</i>)

- Relações típicas:

- ▶ Hiperonímia: espécies, géneros e ordens de plantas e animais

Resultados

Relação	Extraídos		Exemplos
	Total	S/rep	
Hiperonímia TC	24.367	16.228	(<i>desordem,desordem_cerebral</i>) (<i>átomo,átomo_de_carbono</i>)
Hiperonímia	31.254	29.563	(<i>desporto,automobilismo</i>) (<i>estilo_de_música,folk</i>)
Sinonímia	11.872	11.862	(<i>inglês_antigo,anglo-saxão</i>) (<i>estupro,violação</i>)
Parte	1.321	1.287	(<i>jejuno,instestino</i>) (<i>rolas,columbidae</i>)
Finalidade	777	743	(<i>amoxicilina,tratamento_de_infeções</i>) (<i>construção, terracota</i>)
Causador	559	520	(<i>parasita, doença</i>) (<i>doença_neuromuscular,fadiga</i>)

- Relações típicas:

- ▶ Hiperonímia: espécies, géneros e ordens de plantas e animais
- ▶ Sinonímia: variantes portuguesa e brasileira da mesma palavra



Resultados

Relação	Extraídos		Exemplos
	Total	S/rep	
Hiperonímia TC	24.367	16.228	(<i>desordem,desordem_cerebral</i>) (<i>átomo,átomo_de_carbono</i>)
Hiperonímia	31.254	29.563	(<i>desporto,automobilismo</i>) (<i>estilo_de_música,folk</i>)
Sinonímia	11.872	11.862	(<i>inglês_antigo,anglo-saxão</i>) (<i>estupro,violação</i>)
Parte	1.321	1.287	(<i>jejuno,instestino</i>) (<i>rolas,columbidae</i>)
Finalidade	777	743	(<i>amoxicilina,tratamento_de_infeções</i>) (<i>construção, terracota</i>)
Causador	559	520	(<i>parasita, doença</i>) (<i>doença_neuromuscular,fadiga</i>)

- Relações típicas:

- ▶ Hiperonímia: espécies, géneros e ordens de plantas e animais
- ▶ Sinonímia: variantes portuguesa e brasileira da mesma palavra
- ▶ Finalidade: problemas de saúde e suas terapêuticas



Resultados

Relação	Extraídos		Exemplos
	Total	S/rep	
Hiperonímia TC	24.367	16.228	(<i>desordem,desordem_cerebral</i>) (<i>átomo,átomo_de_carbono</i>)
Hiperonímia	31.254	29.563	(<i>desporto,automobilismo</i>) (<i>estilo_de_música,folk</i>)
Sinonímia	11.872	11.862	(<i>inglês_antigo,anglo-saxão</i>) (<i>estupro,violação</i>)
Parte	1.321	1.287	(<i>jejuno,instestino</i>) (<i>rolas,columbidae</i>)
Finalidade	777	743	(<i>amoxicilina,tratamento_de_infeções</i>) (<i>construção, terracota</i>)
Causador	559	520	(<i>parasita, doença</i>) (<i>doença_neuromuscular,fadiga</i>)

- Relações típicas:

- ▶ Hiperonímia: espécies, géneros e ordens de plantas e animais
- ▶ Sinonímia: variantes portuguesa e brasileira da mesma palavra
- ▶ Finalidade: problemas de saúde e suas terapêuticas
- ▶ Causador: causas/efeitos de problemas de saúde



Avaliação manual

- Escala proposta por [Freitas, 2007]:

⁴**Epinecrophylla spodionota** *é uma espécie de ave...*

Avaliação manual

- Escala proposta por [Freitas, 2007]:
 - ▶ 3: Triplo correcto
(*legionelose* SINONIMO_N_DE *doença_do_legionário*)

⁴**Epinecrophylla spodionota** é uma espécie de ave...

Avaliação manual

- Escala proposta por [Freitas, 2007]:
 - ▶ 3: Triplo correcto
(*legionelose* SINONIMO_N_DE *doença_do_legionário*)
 - ▶ 2: Triplo “estranho” por falta de um modificador
(*ave* HIPERONIMO_DE *spodionota*⁴)

⁴**Epinecrophylla spodionota** é uma espécie de ave...

Avaliação manual

- Escala proposta por [Freitas, 2007]:
 - ▶ 3: Triplo correcto
(*legionelose* SINONIMO_N_DE *doença_do legionário*)
 - ▶ 2: Triplo “estranho” por falta de um modificador
(*ave* HIPERONIMO_DE *spodionota*⁴)
 - ▶ 1: Triplo correcto, mas muito genérico ou específico para ser útil
(*região_noroeste* HIPERONIMO_DE *região_noroeste_do Canadá*)

⁴*Epinecrophylla spodionota* é uma espécie de ave...

Avaliação manual

- Escala proposta por [Freitas, 2007]:
 - ▶ 3: Triplo correcto
(*legionelose* SINONIMO_N_DE *doença_do legionário*)
 - ▶ 2: Triplo “estranho” por falta de um modificador
(*ave* HIPERONIMO_DE *spodionota*⁴)
 - ▶ 1: Triplo correcto, mas muito genérico ou específico para ser útil
(*região_noroeste* HIPERONIMO_DE *região_noroeste_do Canadá*)
 - ▶ 0: Triplo errado (*sub-grupos* FINALIDADE_DE *cefetamet*)

⁴**Epinecrophylla spodionota** é uma espécie de ave...

Avaliação manual

- Escala proposta por [Freitas, 2007]:
 - ▶ 3: Triplo correcto
(*legionelose* SINONIMO_N_DE *doença_do_legionário*)
 - ▶ 2: Triplo “estranho” por falta de um modificador
(*ave* HIPERONIMO_DE *spodionota*⁴)
 - ▶ 1: Triplo correcto, mas muito genérico ou específico para ser útil
(*região_noroeste* HIPERONIMO_DE *região_noroeste_do_Canadá*)
 - ▶ 0: Triplo errado (*sub-grupos* FINALIDADE_DE *cefetamet*)

Relação	Avaliados	3(%)	2(%)	1(%)	0(%)	CcEx(%)	CcRel(%)
Hiperonímia TC	323	35,0	4,2	42,1	18,7	57,3	82,7
Hiperonímia	322	57,5	33,8	1,6	7,1	89,8	93,1
Sinonímia	286	85,7	7,3	0,4	6,6	90,0	91,6
Parte	268	44,2	26,7	8,4	20,7	63,1	78,4
Finalidade	264	53,0	16,5	4,0	26,5	71,2	82,2
Causador	244	41,8	24,6	7,8	25,8	61,5	79,5

⁴*Epinecrophylla spodionota* é uma espécie de ave...

Eficiência dos padrões

Relação	Padrão	Extraídos	Avaliados			
			3	2	1	0
Hiperonímia	<i>termo composto</i>	24.367	72	7	75	32
Hiperonímia	é uma espécie de	15.824	54	96	0	0
Hiperonímia	é um uma	10.960	87	11	0	15
Hiperonímia	é um género de	2.402	24	0	0	0
Sinonímia	ou	4.886	154	2	0	2
Sinonímia	também conhecido a os as por como	3.016	60	4	0	4
Parte	inclui incluem	471	34	0	2	15
Parte	grupo de	158	17	3	1	0
Finalidade	utilizado a os as para como em no na	376	71	16	1	20
Finalidade	usado a os as para como em no na	237	41	3	1	4
Causador	causado a os as	165	27	11	1	10



Proposta para validação automática

- Métricas de semelhança distribucional na Web [Bollegala et al., 2007]

$$\text{WebJaccard}(X, Y) = \frac{P(X \cap Y)}{P(X) + P(Y) - P(X, Y)} \quad (1)$$

$$\text{WebOverlap}(X, Y) = \frac{P(X \cap Y)}{\min(P(X), P(Y))} \quad (2)$$

$$\text{WebDice}(X, Y) = \frac{2 * P(X \cap Y)}{P(X) + P(Y)} \quad (3)$$

$$\text{WebPMI}(X, Y) = \log_2 \left(\frac{P(X \cap Y)}{P(X) * P(Y)} * N \right) \quad (4)$$



Padrões utilizados na validação automática

- Seguindo [Oliveira, 2009]
 - ▶ $X = XR, Y = RY$



Padrões utilizados na validação automática

- Seguindo [Oliveira, 2009]
 - ▶ $X = XR, Y = RY$
 - ▶ $X \cap Y = XRY$



Padrões utilizados na validação automática

- Seguindo [Oliveira, 2009]
 - ▶ $X = XR, Y = RY$
 - ▶ $X \cap Y = XRY$

Relação	Padrão indicador (R)
Hiperonímia	é são um uma
Sinonímia	também conhecido conhecida chamado chamada designado designada de por pela
Parte-de	tem possui engloba abrange inclui têm um uma vários alguns
Causa	devido derivado derivada causado causada resultado efeito consequência a ao à por pelo pela de do da
Finalidade	usado usada utilizado utilizada através objectivo finalidade intuito serve no na para de o a um uma

Padrões utilizados na validação automática

- Seguindo [Oliveira, 2009]
 - ▶ $X = XR, Y = RY$
 - ▶ $X \cap Y = XRY$

Relação	Padrão indicador (R)
Hiperonímia	é são um uma
Sinonímia	também conhecido conhecida chamado chamada designado designada de por pela
Parte-de	tem possui engloba abrange inclui têm um uma vários alguns
Causa	devido derivado derivada causado causada resultado efeito consequência a ao à por pelo pela de do da
Finalidade	usado usada utilizado utilizada através objectivo finalidade intuito serve no na para de o a um uma

- Correlações com a avaliação manual < 20%



Comparação com o PAPEL [Gonçalo Oliveira et al., 2010]

- 4.958 termos em ambos

	Tripos		Em ambos		Args em ambos		Um arg em ambos		Novos Wikipédia	
	PAPEL	Wikipédia	Quant	(\approx)	Quant	(\approx)	Quant	(\approx)	Quant	(\approx)
Sinonímia	79,161	11,862	122	1.1%	310	2.6%	3,220	27.1%	8,210	69.2%
Hiperonímia	62,591	45,791	137	0.3%	496	1.1%	28,011	61.2%	17,147	37.4%
Causa	7,966	520	0	0%	61	11.7%	218	41.9%	240	46.2%
Finalidade	8,312	743	0	0%	33	4.4%	279	37.6%	431	58%
Parte.de	13,389	1,287	2	0.2%	19	1.5%	261	20.3%	1,005	78%



Comparação com o PAPEL [Gonçalo Oliveira et al., 2010]

- 4.958 termos em ambos

	Tripos		Em ambos		Args em ambos		Um arg em ambos		Novos Wikipédia	
	PAPEL	Wikipédia	Quant	(\approx)	Quant	(\approx)	Quant	(\approx)	Quant	(\approx)
Sinonímia	79,161	11,862	122	1.1%	310	2.6%	3,220	27.1%	8,210	69.2%
Hiperonímia	62,591	45,791	137	0.3%	496	1.1%	28,011	61.2%	17,147	37.4%
Causa	7,966	520	0	0%	61	11.7%	218	41.9%	240	46.2%
Finalidade	8,312	743	0	0%	33	4.4%	279	37.6%	431	58%
Parte.de	13,389	1,287	2	0.2%	19	1.5%	261	20.3%	1,005	78%

- O conhecimento extraído de enciclopédias pode ser de grande utilidade no aumento de recursos lexicais



Comparação com o PAPEL [Gonçalo Oliveira et al., 2010]

- 4.958 termos em ambos

	Tripos		Em ambos		Args em ambos		Um arg em ambos		Novos Wikipédia	
	PAPEL	Wikipédia	Quant	(\approx)	Quant	(\approx)	Quant	(\approx)	Quant	(\approx)
Sinonímia	79,161	11,862	122	1.1%	310	2.6%	3,220	27.1%	8,210	69.2%
Hiperonímia	62,591	45,791	137	0.3%	496	1.1%	28,011	61.2%	17,147	37.4%
Causa	7,966	520	0	0%	61	11.7%	218	41.9%	240	46.2%
Finalidade	8,312	743	0	0%	33	4.4%	279	37.6%	431	58%
Parte.de	13,389	1,287	2	0.2%	19	1.5%	261	20.3%	1,005	78%

- O conhecimento extraído de enciclopédias pode ser de grande utilidade no aumento de recursos lexicais
 - ▶ O conhecimento em dicionários será mais fulcral



Trabalho futuro

- Várias limitações na nossa abordagem:



Trabalho futuro

- Várias limitações na nossa abordagem:
 - ▶ Explorar outras formas de validação automática



Trabalho futuro

- Várias limitações na nossa abordagem:
 - ▶ Explorar outras formas de validação automática
 - ▶ Apurar a relevância de termos compostos



Trabalho futuro

- Várias limitações na nossa abordagem:
 - ▶ Explorar outras formas de validação automática
 - ▶ Apurar a relevância de termos compostos
 - ▶ Formalismo para gramáticas com dois níveis de análise (átomo + POS)



Trabalho futuro

- Várias limitações na nossa abordagem:
 - ▶ Explorar outras formas de validação automática
 - ▶ Apurar a relevância de termos compostos
 - ▶ Formalismo para gramáticas com dois níveis de análise (átomo + POS)
 - ▶ Abordagem através de aprendizagem levemente supervisionada [Wu and Weld, 2010]



Trabalho futuro

- Várias limitações na nossa abordagem:
 - ▶ Explorar outras formas de validação automática
 - ▶ Apurar a relevância de termos compostos
 - ▶ Formalismo para gramáticas com dois níveis de análise (átomo + POS)
 - ▶ Abordagem através de aprendizagem levemente supervisionada [Wu and Weld, 2010]
 - ▶ Associação automática da informação extraída à ontologia do Onto.PT (como [Ruiz-Casado et al., 2007], continuando o trabalho [Gonçalo Oliveira and Gomes, 2010])



Trabalho futuro

- Várias limitações na nossa abordagem:
 - ▶ Explorar outras formas de validação automática
 - ▶ Apurar a relevância de termos compostos
 - ▶ Formalismo para gramáticas com dois níveis de análise (átomo + POS)
 - ▶ Abordagem através de aprendizagem levemente supervisionada [Wu and Weld, 2010]
 - ▶ Associação automática da informação extraída à ontologia do Onto.PT (como [Ruiz-Casado et al., 2007], continuando o trabalho [Gonçalo Oliveira and Gomes, 2010])
- Em <http://ontopt.dei.uc.pt> disponibilizamos, entre outros recursos, as gramáticas utilizadas



Referências I



Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., and Hellmann, S. (2009).

Dbpedia – a crystallization point for the web of data.

Web Semantics: Science, Services and Agents on the World Wide Web, 7(3):154–165.



Bollegala, D., Matsuo, Y., and Ishizuka, M. (2007).

Measuring semantic similarity between words using web search engines.

In *Proc. 16th International conference on World Wide Web (WWW'07)*, pages 757–766, New York, NY, USA. ACM.



Fellbaum, C., editor (1998).

WordNet: An Electronic Lexical Database (Language, Speech, and Communication).

The MIT Press.



Freitas, M. C. (2007).

Elaboração automática de ontologias de domínio: discussão e resultados.

PhD thesis, Pontifícia Universidade Católica do Rio de Janeiro.



Gonçalo Oliveira, H. and Gomes, P. (2010).

Towards the automatic creation of a wordnet from a term-based lexical network.

In *Proceedings of the ACL Workshop TextGraphs-5: Graph-based Methods for Natural Language Processing*.

No prelo.



Gonçalo Oliveira, H., Santos, D., and Gomes, P. (2010).

Extracção de relações semânticas entre palavras a partir de um dicionário: o PAPEL e sua avaliação.

Linguamática, 2(1):77–93.

Nova versão, revista e aumentada, da publicação Gonçalo Oliveira et al (2009), no STIL 2009.



Referências II



Hirst, G. (2004).

Ontology and the lexicon.

In Staab, S. and Studer, R., editors, *Handbook on Ontologies*, International Handbooks on Information Systems, pages 209–230. Springer.



Oliveira, P. C. (2009).

Probabilistic reasoning in the semantic web using markov logic.

Master's thesis, Universidade de Coimbra, Faculdade de Ciências e Tecnologia, Departamento de Engenharia Informática.



Ruiz-Casado, M., Alfonseca, E., and Castells, P. (2007).

Automatising the learning of lexical patterns: An application to the enrichment of wordnet by extracting semantic relationships from wikipedia.

Data Knowledge Engineering, 61(3):484–499.



Wu, F. and Weld, D. S. (2010).

Open information extraction using wikipedia.

In *Proc. of 48th Annual Meeting of the Association for Computational Linguistics*, pages 118–127, Uppsala, Sweden. Association for Computational Linguistics.



Obrigado!

