Onto.PT: Construção automática de uma ontologia lexical para o português

Hugo Gonçalo Oliveira[1], Paulo Gomes

{hroliv,pgomes}@dei.uc.pt

Cognitive & Media Systems Group
CISUC, Universidade de Coimbra

9 de Dezembro de 2010

[1] financiado pela bolsa FCT SFRH/BD/44955/2008

1. Introdução

2. Fases
   - Extracção de informação
   - Descoberta de conceitos/synsets
   - Juntar recursos baseados em synsets
   - Quantificar triplos
   - Associar termos a synsets
   - Organização do conhecimento

3. Observações finais

Introdução

- É partilhada uma enorme quantidade de informação **escrita**

Introdução

- É partilhada uma enorme quantidade de informação **escrita**
- Manipulação optimizada através de:
  - Recuperação de informação
    - ★ *Que documentos estão relacionados com determinado tema?*

Introdução

- É partilhada uma enorme quantidade de informação **escrita**
- Manipulação optimizada através de:
  - Recuperação de informação
    - ★ *Que documentos estão relacionados com determinado tema?*
  - Sumarização automática
    - ★ *Quais os conteúdos chave de cada documento?*

Introdução

- É partilhada uma enorme quantidade de informação **escrita**
- Manipulação optimizada através de:
  - Recuperação de informação
    - ★ *Que documentos estão relacionados com determinado tema?*
  - Sumarização automática
    - ★ *Quais os conteúdos chave de cada documento?*
  - Resposta automática a outras perguntas
    - ★ *Como se faz X? Quem é o responsável por Y? Quando nasceu Z?*

Introdução

- É partilhada uma enorme quantidade de informação **escrita**
- Manipulação optimizada através de:
  - Recuperação de informação
    - ★ *Que documentos estão relacionados com determinado tema?*
  - Sumarização automática
    - ★ *Quais os conteúdos chave de cada documento?*
  - Resposta automática a outras perguntas
    - ★ *Como se faz X? Quem é o responsável por Y? Quando nasceu Z?*
- Necessário estruturar e **interpretar** a linguagem natural...

Introdução

- É partilhada uma enorme quantidade de informação **escrita**
- Manipulação optimizada através de:
  - Recuperação de informação
    - ★ *Que documentos estão relacionados com determinado tema?*
  - Sumarização automática
    - ★ *Quais os conteúdos chave de cada documento?*
  - Resposta automática a outras perguntas
    - ★ *Como se faz X? Quem é o responsável por Y? Quando nasceu Z?*
- Necessário estruturar e **interpretar** a linguagem natural...
- Melhor acesso a informação léxico-semântica!

Ontologias lexicais

- Definição:
  - Ontologia + léxico [Hirst, 2004]

Ontologias lexicais

- Definição:
  - Ontologia + léxico [Hirst, 2004]
  - Estruturadas em palavras e no seu significado

Ontologias lexicais

- Definição:
  - Ontologia + léxico [Hirst, 2004]
  - Estruturadas em palavras e no seu significado
  - Abranger toda uma língua, sem domínio específico

Ontologias lexicais

- Definição:
  - Ontologia + léxico [Hirst, 2004]
  - Estruturadas em palavras e no seu significado
  - Abranger toda uma língua, sem domínio específico

- Exemplo: Princeton WordNet [Fellbaum, 1998]

- S: (n) **bank** (sloping land (especially the slope beside a body of water)) *"they pulled the canoe up on the bank"; "he sat on the bank of the river and watched the currents"*
  - *direct hyponym* / *full hyponym*
  - *direct hypernym* / *inherited hypernym* / *sister term*
    - S: (n) slope, incline, side (an elevated geological formation) *"he climbed the steep slope"; "the house was built on the side of a mountain"*
  - *derivationally related form*
- S: (n) depository financial institution, **bank**, banking concern, banking company (a financial institution that accepts deposits and channels the money into lending activities) *"he cashed a check at the bank"; "that bank holds the mortgage on my home"*
  - *direct hyponym* / *full hyponym*
  - *member holonym*
  - ***direct hypernym*** / *inherited hypernym* / *sister term*
    - S: (n) financial institution, financial organization, financial organisation (an institution (public or private) that collects funds (from the public or other institutions) and invests them in financial assets)
  - *derivationally related form*
- S: (n) **bank** (a long ridge or pile) *"a huge bank of earth"*
- S: (n) **bank** (an arrangement of similar objects in a row or in tiers) *"he operated a bank of switches"*
- S: (n) **bank** (a supply or stock held in reserve for future use (especially in emergencies))
- S: (n) **bank** (the funds held by a gambling house or the dealer in some gambling games) *"he tried to break the bank at Monte Carlo"*
- S: (n) **bank**, cant, camber (a slope in the turn of a road or track; the outside is higher than the inside in order to reduce the effects of centrifugal force)
- S: (n) savings bank, coin bank, money box, **bank** (a container (usually with a slot in the top) for keeping money at home) *"the coin bank was empty"*
- S: (n) **bank**, bank building (a building in which the business of banking transacted) *"the bank is on the corner of Nassau and Witherspoon"*
- S: (n) **bank** (a flight maneuver; aircraft tips laterally about its longitudinal axis (especially in turning)) *"the plane went into a steep bank"*

Ontologias lexicais (aplicações)

- Ajuda à escrita:
  - *Queria comprar um lindo anel, mas não tinha* **dinheiro**
    Adiei esta oferta especial, porque não tinha **dinheiro**...

Ontologias lexicais (aplicações)

- Ajuda à escrita:
  - *Queria comprar um lindo anel, mas não tinha* **dinheiro**
    Adiei esta oferta especial, porque não tinha **dinheiro**...
- É possível substituir algumas palavras, melhorando a sonoridade e **mantendo o significado**?

Ontologias lexicais (aplicações)

- Ajuda à escrita:
  - *Queria comprar um lindo anel, mas não tinha* **dinheiro**
    Adiei esta oferta especial, porque não tinha **dinheiro**...
- É possível substituir algumas palavras, melhorando a sonoridade e **mantendo o significado**?
  - *Queria comprar um lindo anel, mas não tinha* **pastel**
    Adiei esta oferta especial, porque não tinha **capital**...

Ontologias lexicais (aplicações)

- Ajuda à escrita:
  - *Queria comprar um lindo anel, mas não tinha* **dinheiro**
    Adiei esta oferta especial, porque não tinha **dinheiro**...
- É possível substituir algumas palavras, melhorando a sonoridade e **mantendo o significado**?
  - *Queria comprar um lindo anel, mas não tinha* **pastel**
    Adiei esta oferta especial, porque não tinha **capital**...
- Idealmente...
  - *O meu gato tem quatro rodas motrizes.*
  - *Tenho automóveis, talheres e computadores, entre outros animais.*

Ontologias lexicais (aplicações)

- Ajuda à escrita:
  - *Queria comprar um lindo anel, mas não tinha* **dinheiro**
    Adiei esta oferta especial, porque não tinha **dinheiro**...
- É possível substituir algumas palavras, melhorando a sonoridade e **mantendo o significado**?
  - *Queria comprar um lindo anel, mas não tinha* **pastel**
    Adiei esta oferta especial, porque não tinha **capital**...
- Idealmente...
  - *O meu gato tem quatro rodas motrizes.*
  - *Tenho automóveis, talheres e computadores, entre outros animais.*
    - ★ Morfologia: ok, Sintaxe: ok
    - ★ Semântica: tem a certeza que é isto que pretende?

Ontologias lexicais (cont.)

- Outras tarefas:
  - Determinação de semelhanças
  - Tradução automática
  - Estudos sobre a língua.
  - ...

Ontologias lexicais (cont.)

- Outras tarefas:
  - Determinação de semelhanças
  - Tradução automática
  - Estudos sobre a língua.
  - ...
- Muitas vezes criadas de forma manual...
  - Construção e manutenção muito trabalhosas!

Onto.PT [Gonçalo Oliveira and Gomes, 2010b]

- Construção **automática** de uma ontologia lexical para o português

Onto.PT [Gonçalo Oliveira and Gomes, 2010b]

- Construção **automática** de uma ontologia lexical para o português
- A partir de vários recursos
  - Thesaurus
  - Dicionários/enciclopédias
  - Corpos

Onto.PT [Gonçalo Oliveira and Gomes, 2010b]

- Construção **automática** de uma ontologia lexical para o português
- A partir de vários recursos
  - Thesaurus
  - Dicionários/enciclopédias
  - Corpos
- Modelo Wordnet

Onto.PT [Gonçalo Oliveira and Gomes, 2010b]

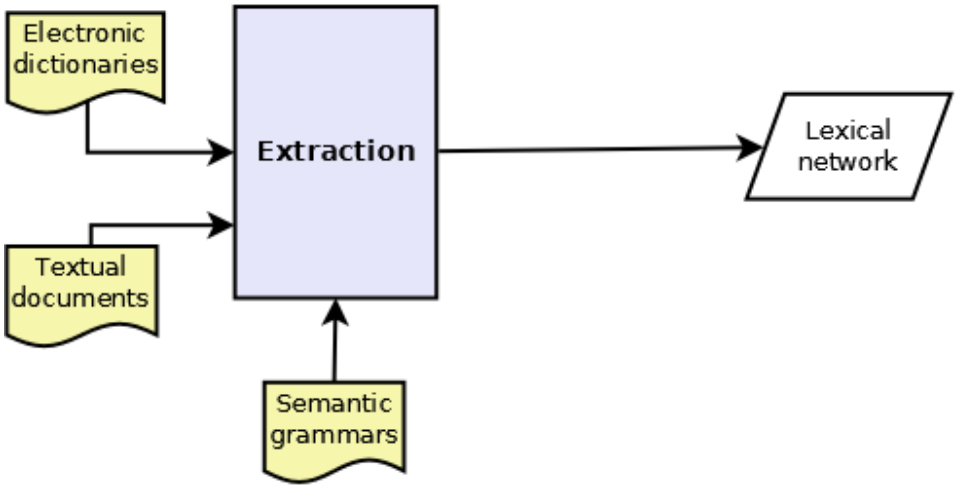
- Construção **automática** de uma ontologia lexical para o português
- A partir de vários recursos
  - Thesaurus
  - Dicionários/enciclopédias
  - Corpos
- Modelo Wordnet
  - *Synsets*: grupos de palavras sinónimas
  - Ligados através de relações semânticas

Onto.PT [Gonçalo Oliveira and Gomes, 2010b]

- Construção **automática** de uma ontologia lexical para o português
- A partir de vários recursos
  - Thesaurus
  - Dicionários/enciclopédias
  - Corpos
- Modelo Wordnet
  - *Synsets*: grupos de palavras sinónimas
  - Ligados através de relações semânticas
- Desambiguação baseada em informação já extraída/sem contexto

Onto.PT [Gonçalo Oliveira and Gomes, 2010b]

- Construção **automática** de uma ontologia lexical para o português
- A partir de vários recursos
  - Thesaurus
  - Dicionários/enciclopédias
  - Corpos
- Modelo Wordnet
  - *Synsets*: grupos de palavras sinónimas
  - Ligados através de relações semânticas
- Desambiguação baseada em informação já extraída/sem contexto
- Explorar métodos de avaliação automática/semi-automática


Electronic dictionaries
Extraction
Lexical network
Textual documents
Semantic grammars

Exemplos

- Dicionários (eg. Dicionário Aberto[2], Wikcionário):
  - `tenreiro, n -- terneiro, novilho ou bezerro.`
    → *terneiro* SINONIMO_DE *tenreiro*
    → *novilho* SINONIMO_DE *tenreiro*
    → *bezerro* SINONIMO_DE *tenreiro*

[2] http://dicionario-aberto.net

Exemplos

- Dicionários (eg. Dicionário Aberto[2], Wikcionário):
  - `tenreiro, n -- terneiro, novilho ou bezerro.`
    → *terneiro* SINONIMO_DE *tenreiro*
    → *novilho* SINONIMO_DE *tenreiro*
    → *bezerro* SINONIMO_DE *tenreiro*
  - `bola, n -- virose que provoca febres e hemorragias`
    → *ébola* CAUSADOR_DE *febres*
    → *ébola* CAUSADOR_DE *hemorragias*

[2] http://dicionario-aberto.net

Exemplos

- Dicionários (eg. Dicionário Aberto[2], Wikcionário):
  - `tenreiro, n -- terneiro, novilho ou bezerro.`
    → *terneiro* SINONIMO_DE *tenreiro*
    → *novilho* SINONIMO_DE *tenreiro*
    → *bezerro* SINONIMO_DE *tenreiro*
  - `bola, n -- virose que provoca febres e hemorragias`
    → *ébola* CAUSADOR_DE *febres*
    → *ébola* CAUSADOR_DE *hemorragias*
- Corpos (eg. artigos da Wikipédia):
  - `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*

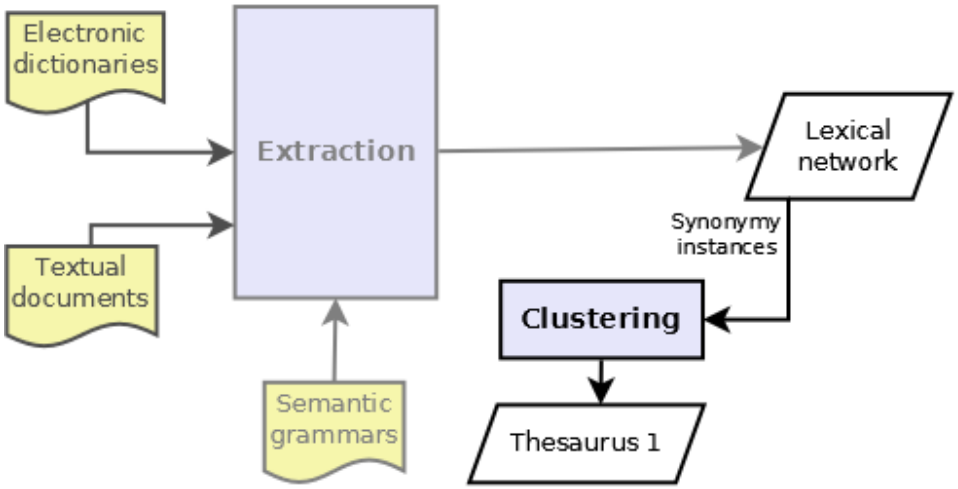
[2] http://dicionario-aberto.net

Triplos extraídos

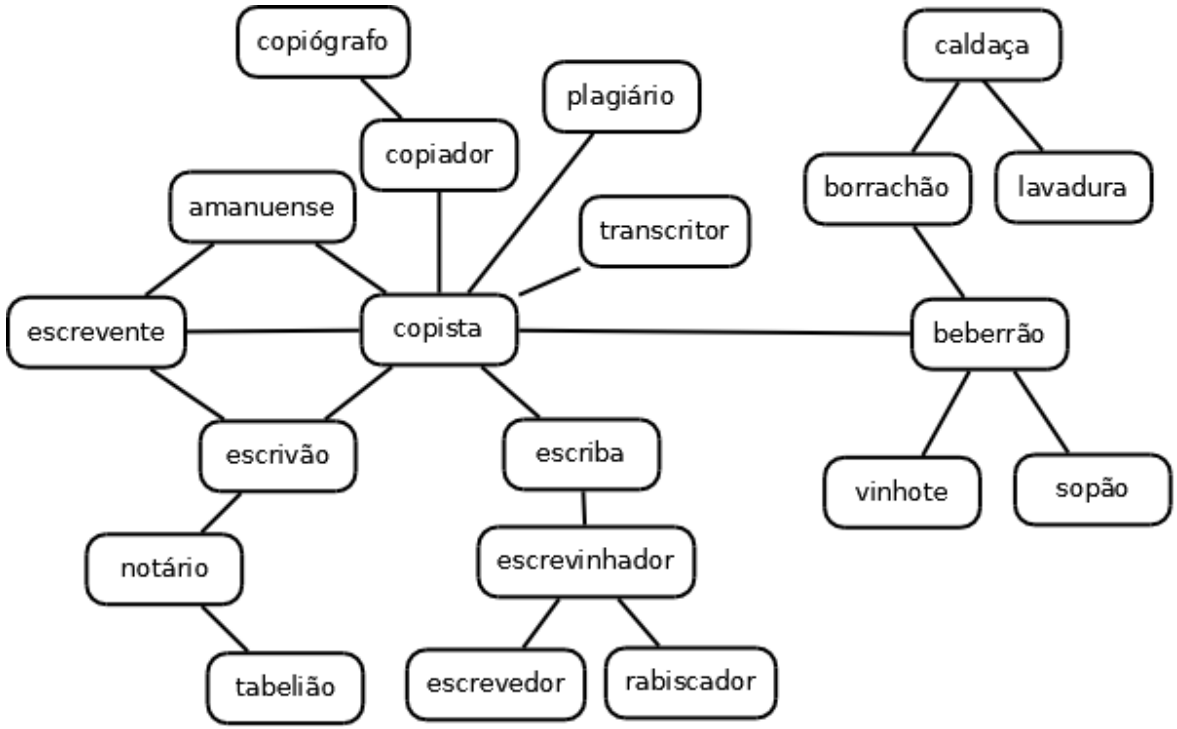
- Dicionário da Língua Portuguesa (PAPEL 2.0[3]) [Gonçalo Oliveira et al., 2010b]
- Dicionário Aberto (DA)
- Resumos da Wikipédia [Gonçalo Oliveira et al., 2010a]

Relação	Argumentos	PAPEL 2.0	DA	Wikipédia	Exemplos
Sinonímia	nome,nome	37,452	20,910	11,862	*auxílio, contributo*
	verbo,verbo	21,465	8,715		*tributar, colectar*
	adj,adj	19,073	7,353		*flexível, moldável*
	adv,adv	1,171	605		*após, seguidamente*
Hiperonímia	nome,nome	62,591	59,887	29,563	*planta, salva*
Parte-de	nome,nome	2,805	1,795	1,287	*cauda, cometa*
	nome,adj	3.721	4,902		*tampa, coberto*
Membro-de	nome,nome	5.929	1,564		*ervilha, Leguminosas*
	adj,nome	883	59		*celular, célula*
Causa	nome,nome	1.013	264	520	*fricção, assadura*
	adj,nome	498	166		*reactivo, reacção*
	verbo,nome	6,399	5,714		*limpar, purgação*
Finalidade	nome,nome	2,886	1,760	743	*defesa, armadura*
	verbo,nome	5,192	3,383		*fazer_rir, comédia*
	verbo,adj	260	186		*corrigir, correccional*

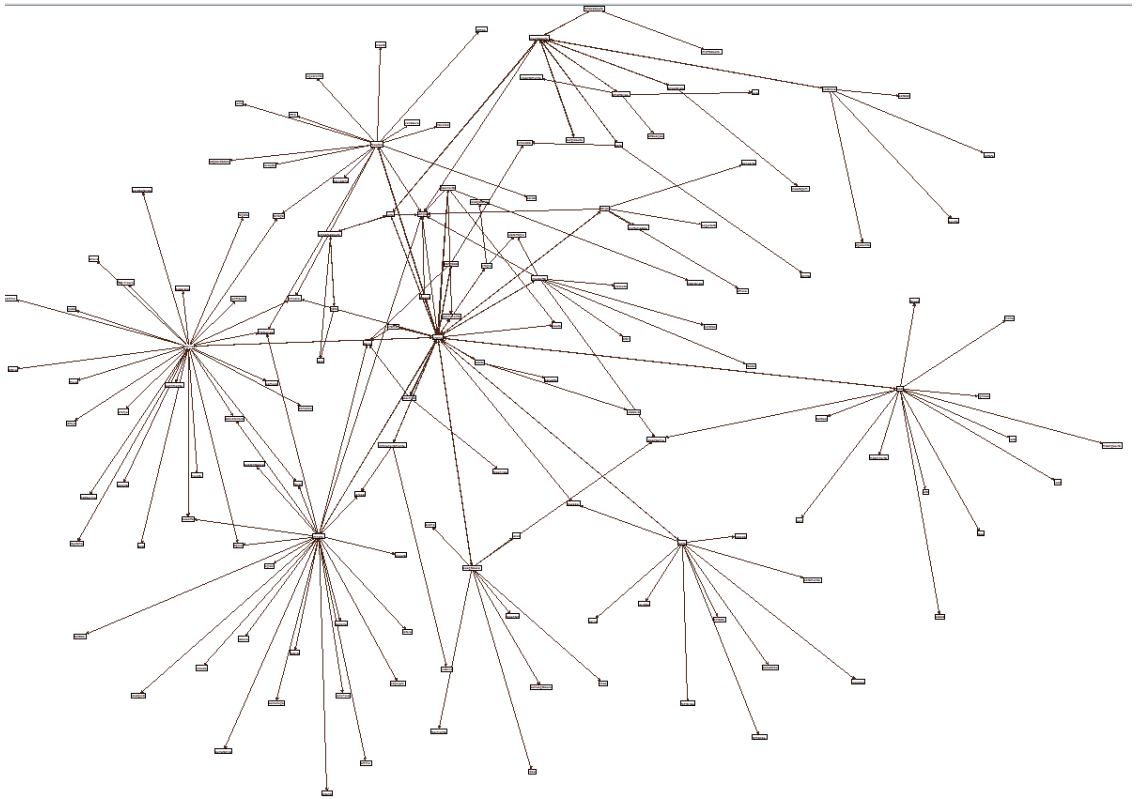
[3] http://www.linguateca.pt/PAPEL


Electronic dictionaries
Extraction
Lexical network
Synonymy instances
Textual documents
Clustering
Semantic grammars
Thesaurus 1

Rede lexical de sinonímia – exemplo

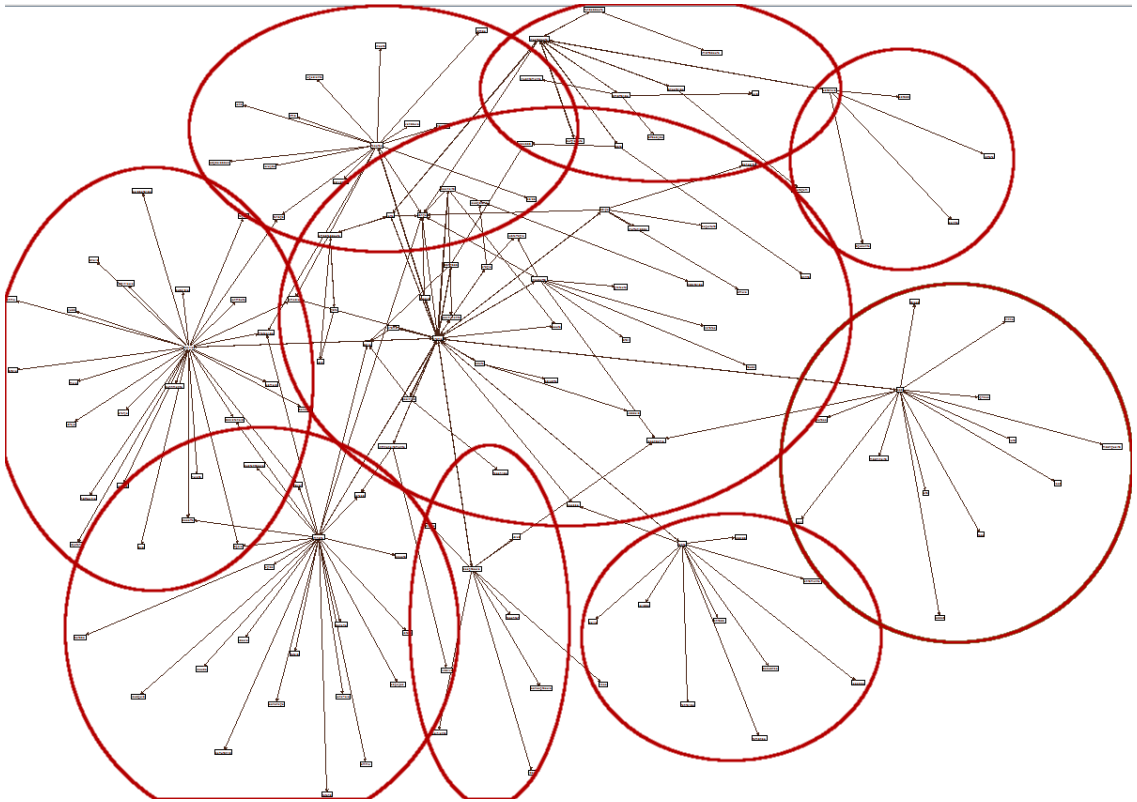


Aglomerados sobressaem em redes de sinonímia...



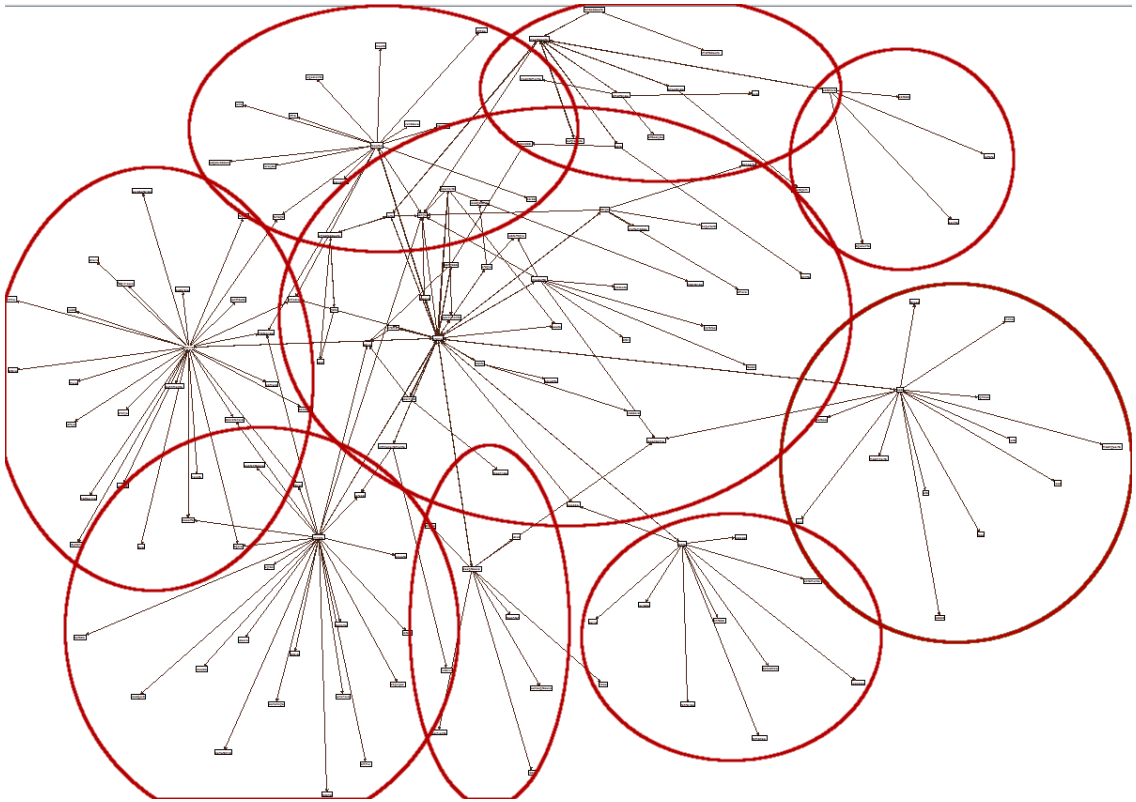
- Objectivo: identificar *synsets* com base em *clusters*

Aglomerados sobressaem em redes de sinonímia...



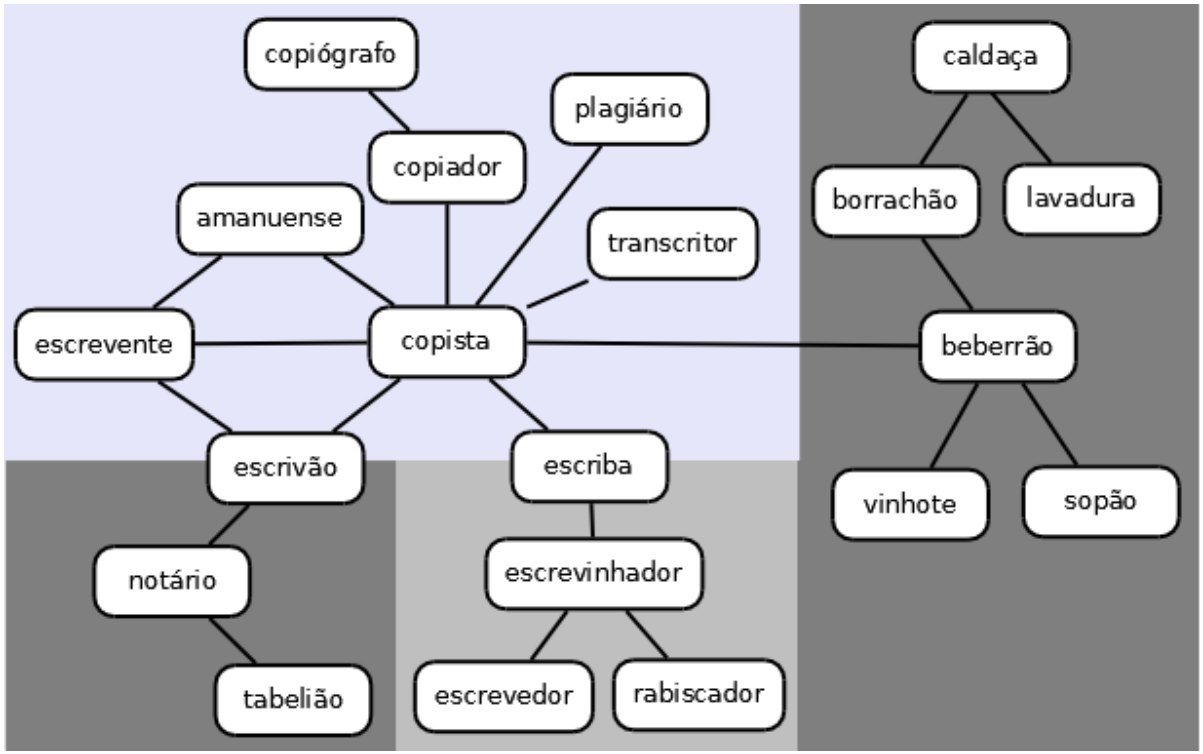
- Objectivo: identificar *synsets* com base em *clusters*
- Abordagem: Algoritmo de *clustering* sobre a rede de sinonímia

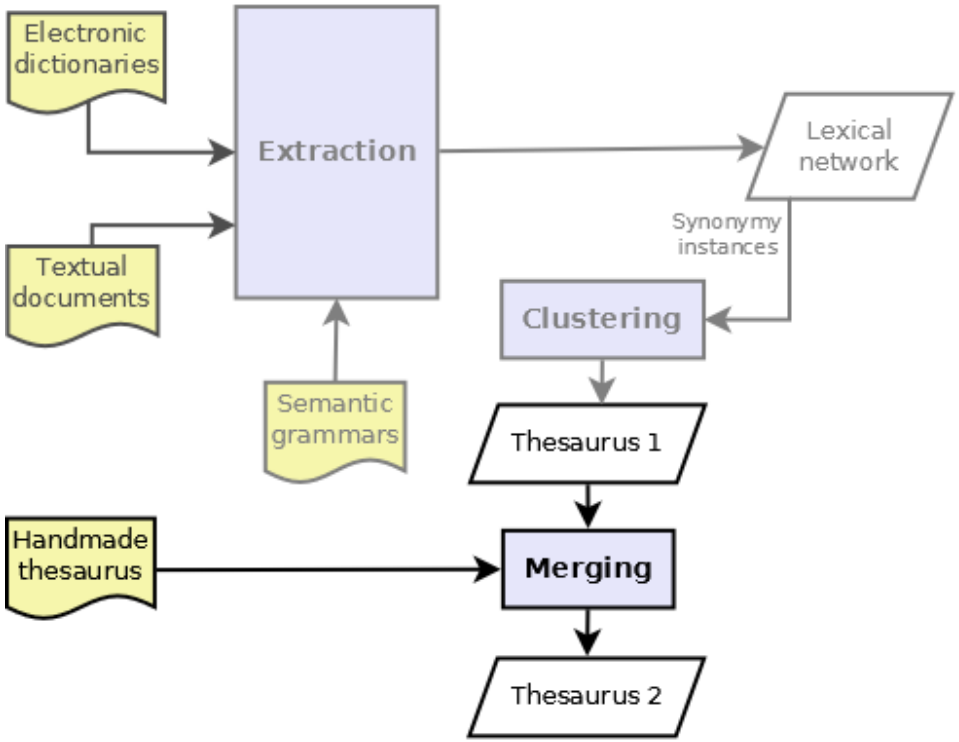
Aglomerados sobressaem em redes de sinonímia...



- Objectivo: identificar *synsets* com base em *clusters*
- Abordagem: Algoritmo de *clustering* sobre a rede de sinonímia
- Manter ambiguidade: os *clusters* podem sobrepor-se!

Clustering – exemplo




Electronic dictionaries
Extraction
Lexical network
Synonymy instances
Textual documents
Clustering
Semantic grammars
Thesaurus 1
Handmade thesaurus
Merging
Thesaurus 2

Juntar *synsets* de diferentes thesaurus

Juntar *synsets* com conteúdo semelhante:

- Por exemplo:
  - $B_1 = (diva, beldade, beleza, deidade, deusa, divindade)$
  - $B_2 = (divindade, deidade, deus, nume)$

Juntar *synsets* de diferentes thesaurus

Juntar *synsets* com conteúdo semelhante:

- Por exemplo:
  - $B_1 = (diva, beldade, beleza, deidade, deusa, divindade)$
  - $B_2 = (divindade, deidade, deus, nume)$
  - $T_1 = (divindade, diva, deusa)$

Juntar *synsets* de diferentes thesaurus

Juntar *synsets* com conteúdo semelhante:

- Por exemplo:
  - $B_1 = (diva, beldade, beleza, deidade, deusa, divindade)$
  - $B_2 = (divindade, deidade, deus, nume)$
  - $T_1 = (divindade, diva, deusa)$
  - $B_1 = B_1 \cup T_1 = (diva, beldade, beleza, deidade, deusa, divindade)$

Primeiros resultados [Gonçalo Oliveira and Gomes, 2010a]

- TeP[4] thesaurus
- OpenThesaurus.PT[5] (OT)

[4] http://www.nilc.icmc.usp.br/tep2/index.htm
[5] http://openthesaurus.caixamagica.pt/

Primeiros resultados [Gonçalo Oliveira and Gomes, 2010a]

- TeP[4] thesaurus
- OpenThesaurus.PT[5] (OT)
- Clusters do PAPEL (CLIP)

[4] http://www.nilc.icmc.usp.br/tep2/index.htm
[5] http://openthesaurus.caixamagica.pt/

Primeiros resultados [Gonçalo Oliveira and Gomes, 2010a]

- TeP[4] thesaurus
- OpenThesaurus.PT[5] (OT)
- Clusters do PAPEL (CLIP)
- TeP + OT + CLIP (TOP)

[4] http://www.nilc.icmc.usp.br/tep2/index.htm
[5] http://openthesaurus.caixamagica.pt/

Primeiros resultados [Gonçalo Oliveira and Gomes, 2010a]

- TeP[4] thesaurus
- OpenThesaurus.PT[5] (OT)
- Clusters do PAPEL (CLIP)
- TeP + OT + CLIP (TOP)

		TeP	OT	CLIP	TOP
Termos	**Quantidade**	17,158	5,819	23,741	30,554
	Ambíguos	5,867	442	12,196	13,294
	Mais ambíguo	20	4	47	21
Synsets	**Quantidade**	8,254	1,872	7,468	9,960
	Tamanho médio	3.51	3.37	12.57	6.6
	Maior	21	14	103	277

Tabela: Thesaurus (de nomes) em números.

[4] http://www.nilc.icmc.usp.br/tep2/index.htm
[5] http://openthesaurus.caixamagica.pt/

Validação manual

	Amostra	Correctos	Incorrectos	N/A	Concordância
CLIP	519 sets	65.8%	31.7%	2.5%	76.1%
CLIP'	310 sets	81.1%	16.9%	2.0%	84.2%
TOP	480 sets	83.2%	15.8%	1.0%	82.3%
TOP'	448 sets	86.8%	12.3%	0.9%	83.0%

Tabela: Resultados da validação manual de synsets.

- CLIP' e TOP' consideram apenas synsets com 10 ou menos termos.

Validação manual

	Amostra	Correctos	Incorrectos	N/A	Concordância
CLIP	519 sets	65.8%	31.7%	2.5%	76.1%
CLIP'	310 sets	81.1%	16.9%	2.0%	84.2%
TOP	480 sets	83.2%	15.8%	1.0%	82.3%
TOP'	448 sets	86.8%	12.3%	0.9%	83.0%

Tabela: Resultados da validação manual de synsets.

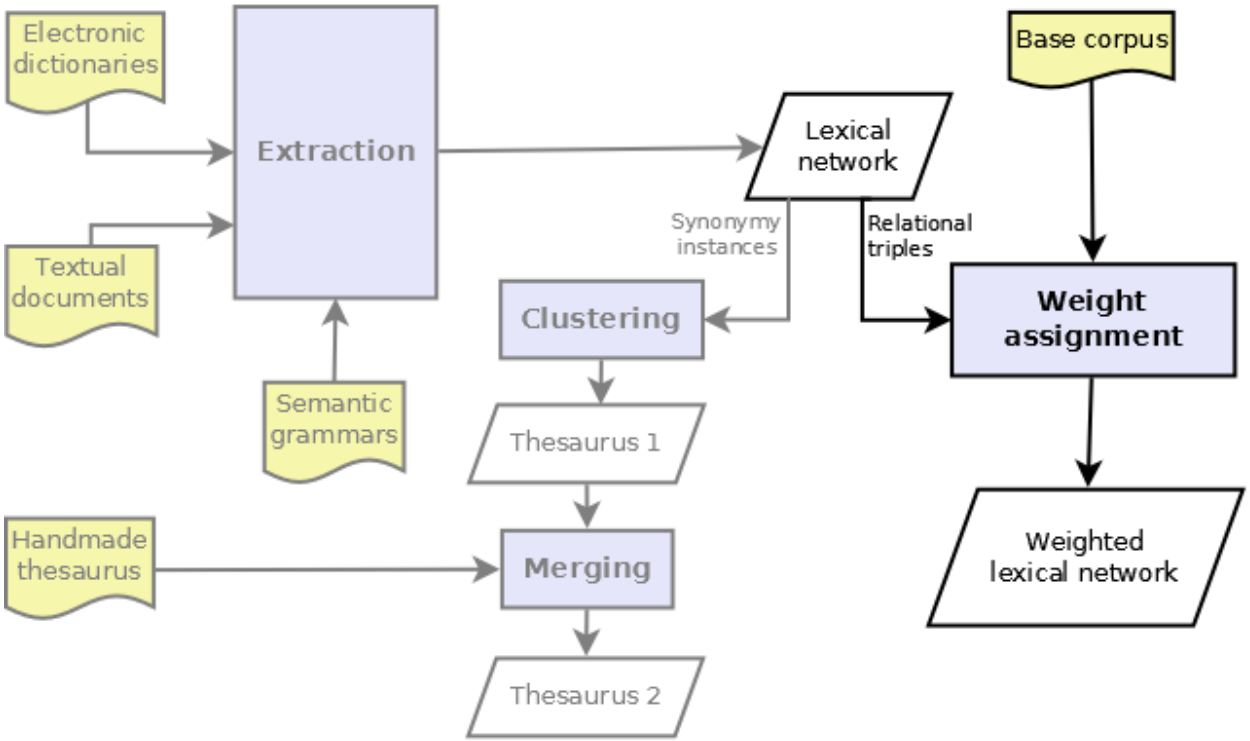
- CLIP' e TOP' consideram apenas synsets com 10 ou menos termos.
  - Melhor qualidade para synsets mais pequenos

Validação manual

	Amostra	Correctos	Incorrectos	N/A	Concordância
CLIP	519 sets	65.8%	31.7%	2.5%	76.1%
CLIP'	310 sets	81.1%	16.9%	2.0%	84.2%
TOP	480 sets	83.2%	15.8%	1.0%	82.3%
TOP'	448 sets	86.8%	12.3%	0.9%	83.0%

Tabela: Resultados da validação manual de synsets.

- CLIP' e TOP' consideram apenas synsets com 10 ou menos termos.
  - Melhor qualidade para synsets mais pequenos
  - Synsets em que uma palavra une dois conceitos, como *excesso* em:
    - ★ *insobriedade, desmedida, imoderação,* **excesso**, *nimiedade, desmando, desbragamento, troco, descontrolo, superabundância, desbunda, desregramento, demasia, incontinência, imodicidade, superação, intemperança, descomedimento, superfluidade, sobejidão, acrasia*


Electronic dictionaries
Textual documents
Semantic grammars
Extraction
Lexical network
Base corpus
Synonymy instances
Relational triples
Clustering
Weight assignment
Thesaurus 1
Handmade thesaurus
Merging
Weighted lexical network
Thesaurus 2

Atribuir pesos com base em...

- Frequência de extracção

Atribuir pesos com base em...

- Frequência de extracção
- Métricas de semelhança distribucional em corpos (e.g. LSA [Deerwester et al., 1990], PMI [Turney, 2001])

Atribuir pesos com base em...

- Frequência de extracção
- Métricas de semelhança distribucional em corpos (e.g. LSA [Deerwester et al., 1990], PMI [Turney, 2001])
- Métricas de semelhança distribucional na Web (e.g. WebJaccard, WebOverlap [Bollegala et al., 2007])

Atribuir pesos com base em...

- Frequência de extracção
- Métricas de semelhança distribucional em corpos (e.g. LSA [Deerwester et al., 1990], PMI [Turney, 2001])
- Métricas de semelhança distribucional na Web (e.g. WebJaccard, WebOverlap [Bollegala et al., 2007])

- Algumas conclusões:
  - Correlações elevadas ($\approx 58\%$) entre a correcção de triplos de hiperonímia e os valores do LSA para os termos envolvidos [Costa et al., 2010]

Atribuir pesos com base em...

- Frequência de extracção
- Métricas de semelhança distribucional em corpos (e.g. LSA [Deerwester et al., 1990], PMI [Turney, 2001])
- Métricas de semelhança distribucional na Web (e.g. WebJaccard, WebOverlap [Bollegala et al., 2007])

- Algumas conclusões:
  - Correlações elevadas ($\approx 58\%$) entre a correcção de triplos de hiperonímia e os valores do LSA para os termos envolvidos [Costa et al., 2010]
  - Métricas de semelhança conjugadas com padrões indicadores podem ser utilizadas para validar triplos [Costa et al., 2011]

Validação automática

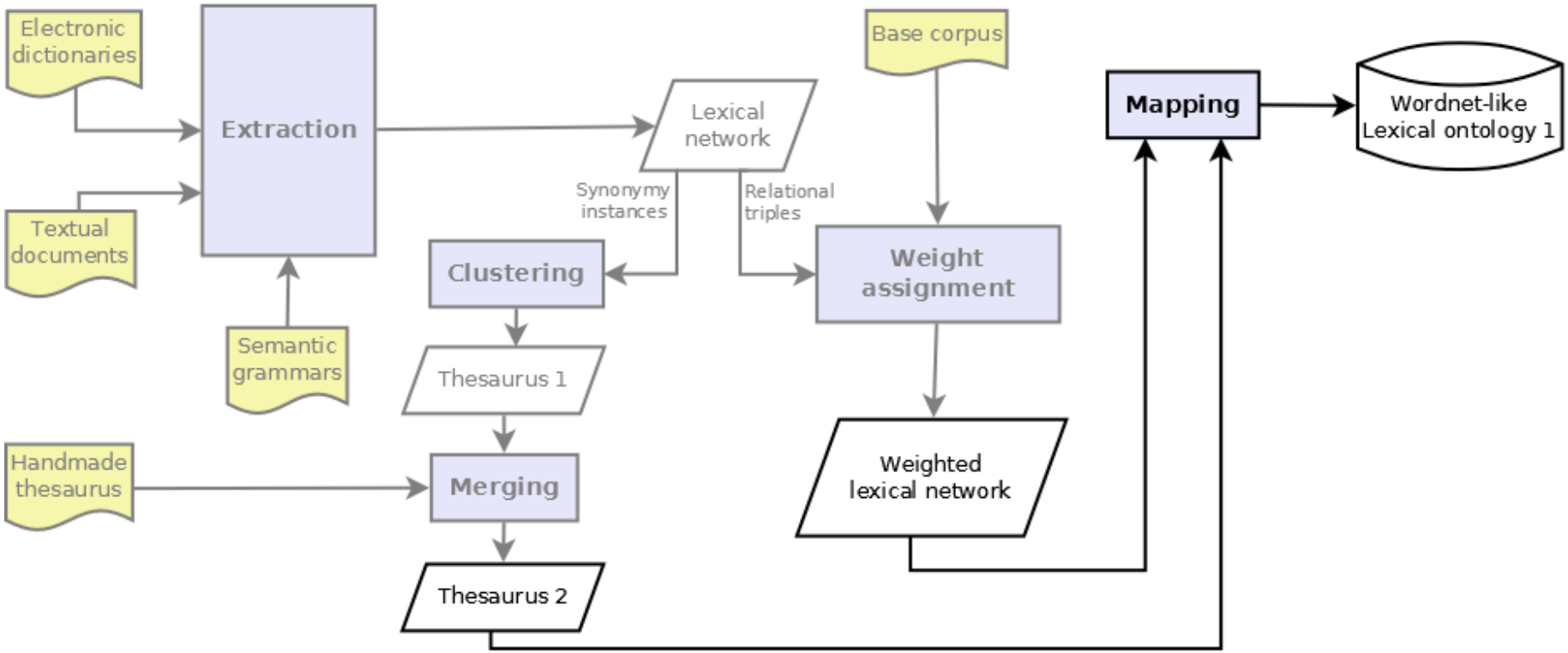
- *animal* HIPERONIMO_DE *cão*
- *porta* PARTE_DE *carro*

Validação automática

- *animal* HIPERONIMO_DE *cão*
  - *animais* `[tal|tais]*` como *cães*
  - *cão* é `um|uma [tipo|forma|variedade|... de]*` *animal*
  - ...
- *porta* PARTE_DE *carro*

Validação automática

- *animal* HIPERONIMO_DE *cão*
  - *animais* `[tal|tais]*` como *cães*
  - *cão* é `um|uma [tipo|forma|variedade|... de]*` *animal*
  - ...
- *porta* PARTE_DE *carro*
  - *porta* `[é uma parte]* de|da|do]` *carro*
  - *carro* `[tem|possui] [um |uma]*` *porta*
  - *carro* `[é formado|constituído|... por [um|uma]*` *porta*
  - ...


Electronic dictionaries
Textual documents
Semantic grammars
Extraction
Lexical network
Synonymy instances
Relational triples
Base corpus
Clustering
Weight assignment
Thesaurus 1
Handmade thesaurus
Merging
Weighted lexical network
Thesaurus 2
Mapping
Wordnet-like Lexical ontology 1

Objectivo

- Entrada:
  - Thesaurus T, que contém *synsets*
  - Rede lexical, N, com termos nos nós e arcos com um tipo R

Objectivo

- Entrada:
  - Thesaurus T, que contém *synsets*
  - Rede lexical, N, com termos nos nós e arcos com um tipo R
- Objectivo: passar de $a\ R\ b \in N$ para $A\ R\ B$, $(A, B) \in T$

Objectivo

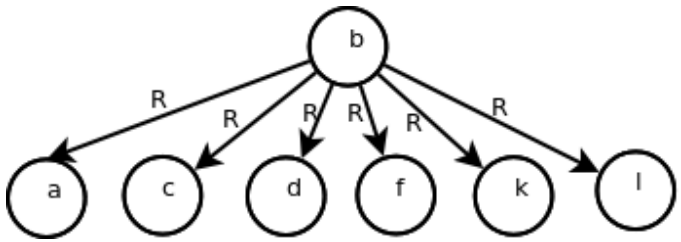
- Entrada:
  - Thesaurus T, que contém *synsets*
  - Rede lexical, N, com termos nos nós e arcos com um tipo R
- Objectivo: passar de $a\ R\ b \in N$ para $A\ R\ B$, $(A, B) \in T$
  - *porta* PARTE_DE *carro* $\rightarrow$ (*porta*, *entrada*, *portão*) PARTE_DE (*carro*, *automóvel*)

Objectivo

- Entrada:
  - Thesaurus T, que contém *synsets*
  - Rede lexical, N, com termos nos nós e arcos com um tipo R
- Objectivo: passar de $a\ R\ b \in N$ para $A\ R\ B$, $(A, B) \in T$
  - *porta* PARTE_DE *carro* $\rightarrow$ (*porta*, *entrada*, *portão*) PARTE_DE (*carro*, *automóvel*)
- Saída: rede semântica W, cujos nós são *synsets*, que se relacionam entre si por meio de relações semânticas (*wordnet*)

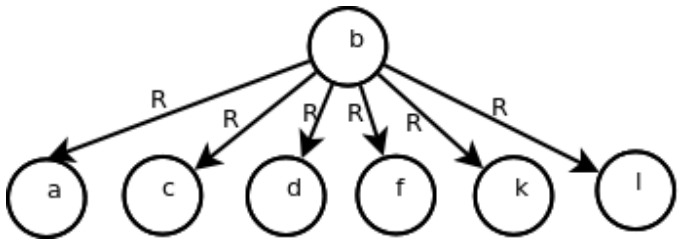
Procedimento exemplo [Gonçalo Oliveira and Gomes, 2010c]

- Olhando para a rede lexical, associar *a* (em *a R b*) a *A*:
  1. Fixar *b*



Procedimento exemplo [Gonçalo Oliveira and Gomes, 2010c]

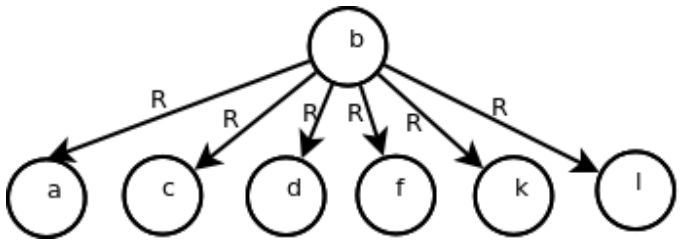
- Olhando para a rede lexical, associar a (em $a\ R\ b$) a A:
  1. Fixar b



  2. $S_a \subset T : S_{ai} \in S_a, a \in S_{ai}$

Procedimento exemplo [Gonçalo Oliveira and Gomes, 2010c]

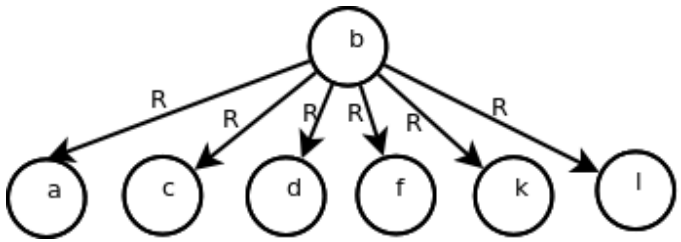
- Olhando para a rede lexical, associar a (em $a\ R\ b$) a A:
  1. Fixar b



  2. $S_a \subset T : S_{ai} \in S_a, a \in S_{ai}$
  3. Para cada $S_{ai} \in S_a$

Procedimento exemplo [Gonçalo Oliveira and Gomes, 2010c]

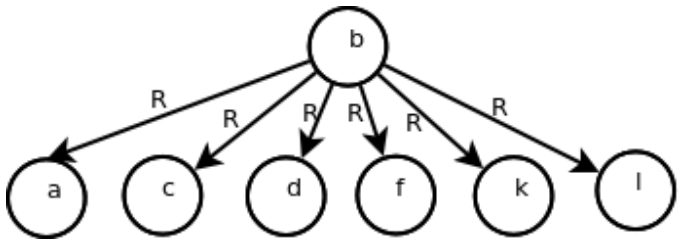
- Olhando para a rede lexical, associar a (em $a\ R\ b$) a A:
  1. Fixar b



  2. $S_a \subset T : S_{ai} \in S_a, a \in S_{ai}$
  3. Para cada $S_{ai} \in S_a$
     - ★ $S_{a1} = (\mathbf{a}, \mathbf{c}, \mathbf{d}, e)$, $p_{a1} = \frac{3}{4}$
     - ★ $S_{a2} = (\mathbf{a}, \mathbf{f}, g)$, $p_{a2} = \frac{2}{3}$
     - ★ $S_{a3} = (\mathbf{a}, h, i, j)$, $p_{a3} = \frac{1}{4}$

Procedimento exemplo [Gonçalo Oliveira and Gomes, 2010c]

- Olhando para a rede lexical, associar a (em $a\ R\ b$) a A:
  1. Fixar b



  2. $S_a \subset T : S_{ai} \in S_a, a \in S_{ai}$
  3. Para cada $S_{ai} \in S_a$
     - ★ $S_{a1} = (\mathbf{a}, \mathbf{c}, \mathbf{d}, e)$, $p_{a1} = \frac{3}{4}$
     - ★ $S_{a2} = (\mathbf{a}, \mathbf{f}, g)$, $p_{a2} = \frac{2}{3}$
     - ★ $S_{a3} = (\mathbf{a}, h, i, j)$, $p_{a3} = \frac{1}{4}$
  4. $a \rightarrow S_{a1}$

Validação

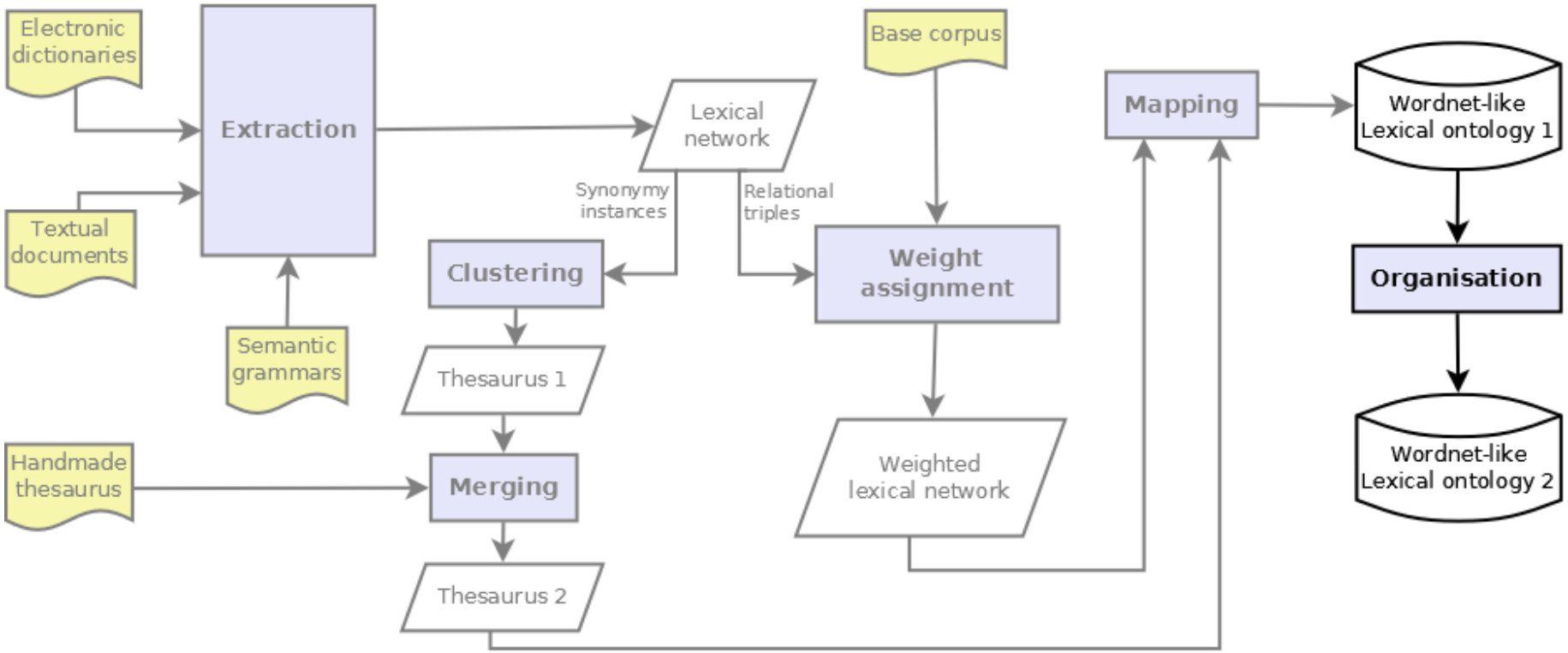
- Automática
  - A partir de ocorrências dos triplos ($A\ R\ B$) em corpos
    1. Compilar um conjunto de padrões indicadores de R, P_i:
    2. Quantificar com ocorrências de $a\ P_i\ b$, $a \in A$, $b \in B$

Validação

- Automática
  - A partir de ocorrências dos triplos ($A\ R\ B$) em corpos
    1. Compilar um conjunto de padrões indicadores de R, P_i:
    2. Quantificar com ocorrências de $a\ P_i\ b$, $a \in A$, $b \in B$
- Semi-automática
  - Mapeamento manual utilizado como recurso dourado
    - ★ Trabalhoso e nem sempre consensual...

Validação

- Automática
  - A partir de ocorrências dos triplos ($A\ R\ B$) em corpos
    1. Compilar um conjunto de padrões indicadores de R, P_i:
    2. Quantificar com ocorrências de $a\ P_i\ b$, $a \in A$, $b \in B$
- Semi-automática
  - Mapeamento manual utilizado como recurso dourado
    - ★ Trabalhoso e nem sempre consensual...
- Manual
  - Julgamentos humanos a ocorrências dos triplos em corpos


Electronic dictionaries
Textual documents
Semantic grammars
Extraction
Lexical network
Synonymy instances
Relational triples
Base corpus
Clustering
Weight assignment
Thesaurus 1
Handmade thesaurus
Merging
Weighted lexical network
Thesaurus 2
Mapping
Wordnet-like Lexical ontology 1
Organisation
Wordnet-like Lexical ontology 2

Organização do conhecimento

- Regras:
  - Transitividade
    - ⋆ se R for transitiva (p.e. Sinonímia, Hiperonímia, ...):
      $(A\ R\ B) \wedge (B\ R\ C) \rightarrow (A\ R\ C)$

[6] http://www.w3.org/2006/03/wn/wn20/

Organização do conhecimento

- Regras:
  - Transitividade
    - ★ se R for transitiva (p.e. Sinonímia, Hiperonímia, ...):
      $(A\ R\ B) \wedge (B\ R\ C) \rightarrow (A\ R\ C)$
  - Herança
    - ★ se R não for uma relação de Hiperonímia ou Hiponímia:
      $(A\ \text{HIPERONIMO\_DE}\ B) \wedge (A\ R\ C) \rightarrow (B\ R\ C)$

[6] http://www.w3.org/2006/03/wn/wn20/

Organização do conhecimento

- Regras:
  - Transitividade
    - ★ se R for transitiva (p.e. Sinonímia, Hiperonímia, ...):
      $(A\ R\ B) \wedge (B\ R\ C) \rightarrow (A\ R\ C)$
  - Herança
    - ★ se R não for uma relação de Hiperonímia ou Hiponímia:
      $(A$ HIPERONIMO_DE $B) \wedge (A\ R\ C) \rightarrow (B\ R\ C)$
  - Auto-hiponímia
    - ★ $(A$ HIPERONIMO_DE $B) \wedge (B$ HIPERONIMO_DE $A) \rightarrow (A \cup B)$

[6] http://www.w3.org/2006/03/wn/wn20/

Organização do conhecimento

- Regras:
  - Transitividade
    - ⋆ se R for transitiva (p.e. Sinonímia, Hiperonímia, ...):
      $(A\ R\ B) \wedge (B\ R\ C) \rightarrow (A\ R\ C)$
  - Herança
    - ⋆ se R não for uma relação de Hiperonímia ou Hiponímia:
      $(A\ \text{HIPERONIMO\_DE}\ B) \wedge (A\ R\ C) \rightarrow (B\ R\ C)$
  - Auto-hiponímia
    - ⋆ $(A\ \text{HIPERONIMO\_DE}\ B) \wedge (B\ \text{HIPERONIMO\_DE}\ A) \rightarrow (A \cup B)$
- Representação num modelo RDF/OWL, semelhante à WordNet-RDF[6]

[6] http://www.w3.org/2006/03/wn/wn20/

Observações finais

- Resposta a:
  - Crescente número de aplicações que necessitam de interpretar a linguagem natural

Observações finais

- Resposta a:
  - Crescente número de aplicações que necessitam de interpretar a linguagem natural
  - Falta de recursos lexico-semânticos **livres** para o português

Observações finais

- Resposta a:
  - Crescente número de aplicações que necessitam de interpretar a linguagem natural
  - Falta de recursos lexico-semânticos **livres** para o português
- Actualizações e recursos disponíveis em `http://ontopt.dei.uc.pt`

Referências I

[Bollegala et al., 2007] 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.

[Costa et al., 2010] Costa, H., Gonçalo Oliveira, H., and Gomes, P. (2010).
The impact of distributional metrics in the quality of relational triples.
In *Proc. ECAI Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2010)*.

[Costa et al., 2011] Costa, H., Gonçalo Oliveira, H., and Gomes, P. (2011).
Using the web to validate semantic relations.
Submitted to ECIR 2011.

[Deerwester et al., 1990] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990).
Indexing by latent semantic analysis.
Journal of the American Society for Information Science, 41:391–407.

[Fellbaum, 1998] Fellbaum, C., editor (1998).
WordNet: An Electronic Lexical Database (Language, Speech, and Communication).
The MIT Press.

[Gonçalo Oliveira et al., 2010a] Gonçalo Oliveira, H., Costa, H., and Gomes, P. (2010a).
Extracção de conhecimento léxico-semântico a partir de resumos da Wikipédia.
In *Proceedings of INFORUM 2010*.

[Gonçalo Oliveira and Gomes, 2010a] Gonçalo Oliveira, H. and Gomes, P. (2010a).
Automatic creation of a conceptual base for portuguese using clustering techniques.
In *Proc. 19th European Conference on Artifical Intelligence (ECAI 2010)*.

[Gonçalo Oliveira and Gomes, 2010b] Gonçalo Oliveira, H. and Gomes, P. (2010b).
Onto.PT: Automatic Construction of a Lexical Ontology for Portuguese.
In *Proceedings of 5th European Starting AI Researcher Symposium (STAIRS 2010)*.

Referências II

[Gonçalo Oliveira and Gomes, 2010c] Gonçalo Oliveira, H. and Gomes, P. (2010c).
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*.

[Gonçalo Oliveira et al., 2010b] Gonçalo Oliveira, H., Santos, D., and Gomes, P. (2010b).
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.

[Hirst, 2004] 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.

[Turney, 2001] Turney, P. D. (2001).
Mining the web for synonyms: PMI–IR versus LSA on TOEFL.
In *Proc. 12th European Conference on Machine Learning (ECML-2001)*, volume 2167, pages 491–502. Springer.

Obrigado!