Onto.PT: Automatic construction of a Lexical Ontology for Portuguese

Hugo Gonçalo Oliveira[1], Paulo Gomes

{hroliv,pgomes}@dei.uc.pt

Cognitive & Media Systems Group
CISUC, University of Coimbra

Lisbon, August 16, 2010

[1] supported by FCT scholarship grant SFRH/BD/44955/2008

Today's applications

- Need to understand information conveyed by natural language

Today's applications

- Need to understand information conveyed by natural language
- Therefore, demand better access to knowledge on words and their meanings!

Today's applications

- Need to understand information conveyed by natural language
- Therefore, demand better access to knowledge on words and their meanings!
- Encoded in lexical ontologies

Lexical ontologies

- Such as Princeton WordNet [Fellbaum, 1998]

Lexical ontologies

- Such as Princeton WordNet [Fellbaum, 1998]
  - Ontology + lexicon [Hirst, 2004]

Lexical ontologies

- Such as Princeton WordNet [Fellbaum, 1998]
  - Ontology + lexicon [Hirst, 2004]
  - Knowledge structured on words and their meanings

Lexical ontologies

- Such as Princeton WordNet [Fellbaum, 1998]
  - Ontology + lexicon [Hirst, 2004]
  - Knowledge structured on words and their meanings
  - Cover the whole language
  - Not based on a specific domain

Lexical ontologies

- Such as Princeton WordNet [Fellbaum, 1998]
  - Ontology + lexicon [Hirst, 2004]
  - Knowledge structured on words and their meanings
  - Cover the whole language
  - Not based on a specific domain
- Typically handcrafted...
  - Construction and maintenance involve time-consuming human effort!

Onto.PT

- Automatic construction of a lexical ontology for Portuguese

Onto.PT

- Automatic construction of a lexical ontology for Portuguese
- Extracted from different sources

Onto.PT

- Automatic construction of a lexical ontology for Portuguese
- Extracted from different sources
  - Manually created thesauri
  - Language dictionaries/encyclopedias
  - Corpora

Onto.PT

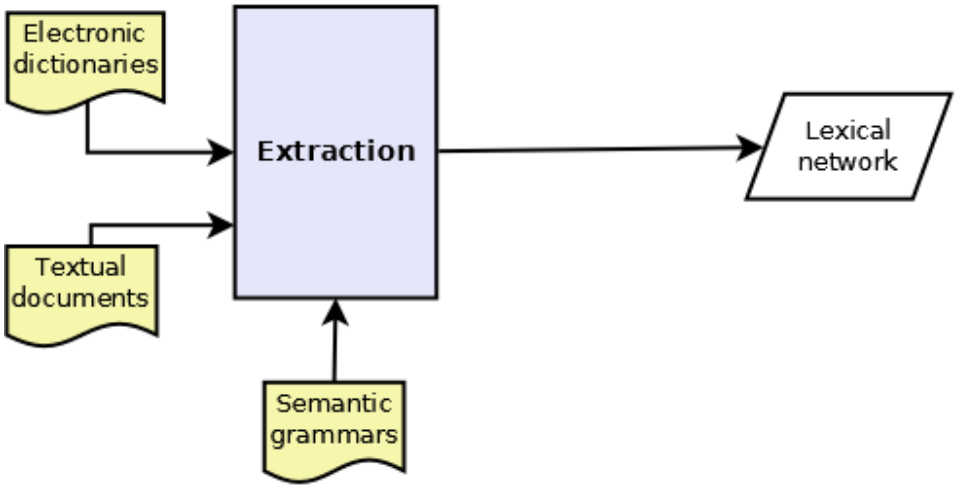
- Automatic construction of a lexical ontology for Portuguese
- Extracted from different sources
  - Manually created thesauri
  - Language dictionaries/encyclopedias
  - Corpora
- Modelled after Princeton WordNet

Onto.PT

- Automatic construction of a lexical ontology for Portuguese
- Extracted from different sources
  - Manually created thesauri
  - Language dictionaries/encyclopedias
  - Corpora
- Modelled after Princeton WordNet
  - Synsets: groups of synonymous words
  - Synset-based relational triples

Onto.PT

- Automatic construction of a lexical ontology for Portuguese
- Extracted from different sources
  - Manually created thesauri
  - Language dictionaries/encyclopedias
  - Corpora
- Modelled after Princeton WordNet
  - Synsets: groups of synonymous words
  - Synset-based relational triples
- WSD based on the knowledge already extracted, not on the context


Electronic dictionaries
Extraction
Lexical network
Textual documents
Semantic grammars

Examples

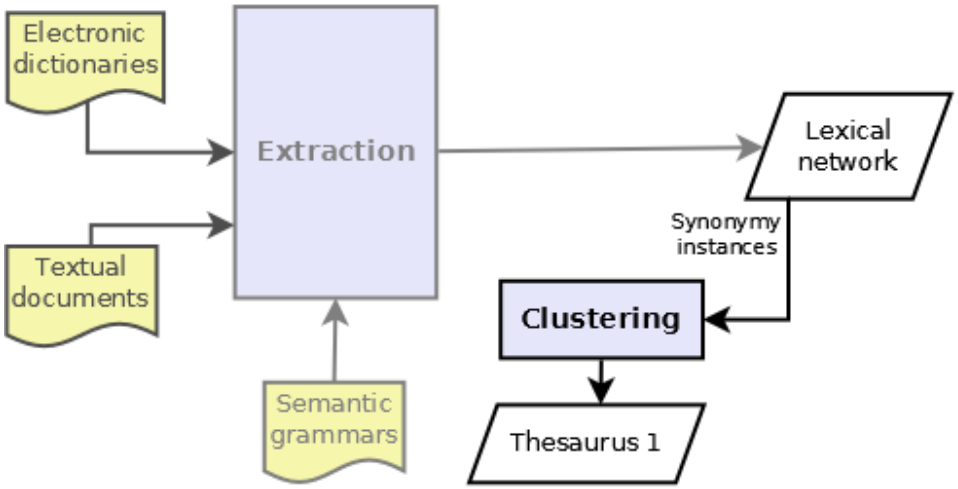
- From dictionaries:
  - `tenreiro, n -- terneiro, novilho ou bezerro.`
    → *terneiro* SYNONYM_OF *tenreiro*
    → *novilho* SYNONYM_OF *tenreiro*
    → *bezerro* SYNONYM_OF *tenreiro*

Examples

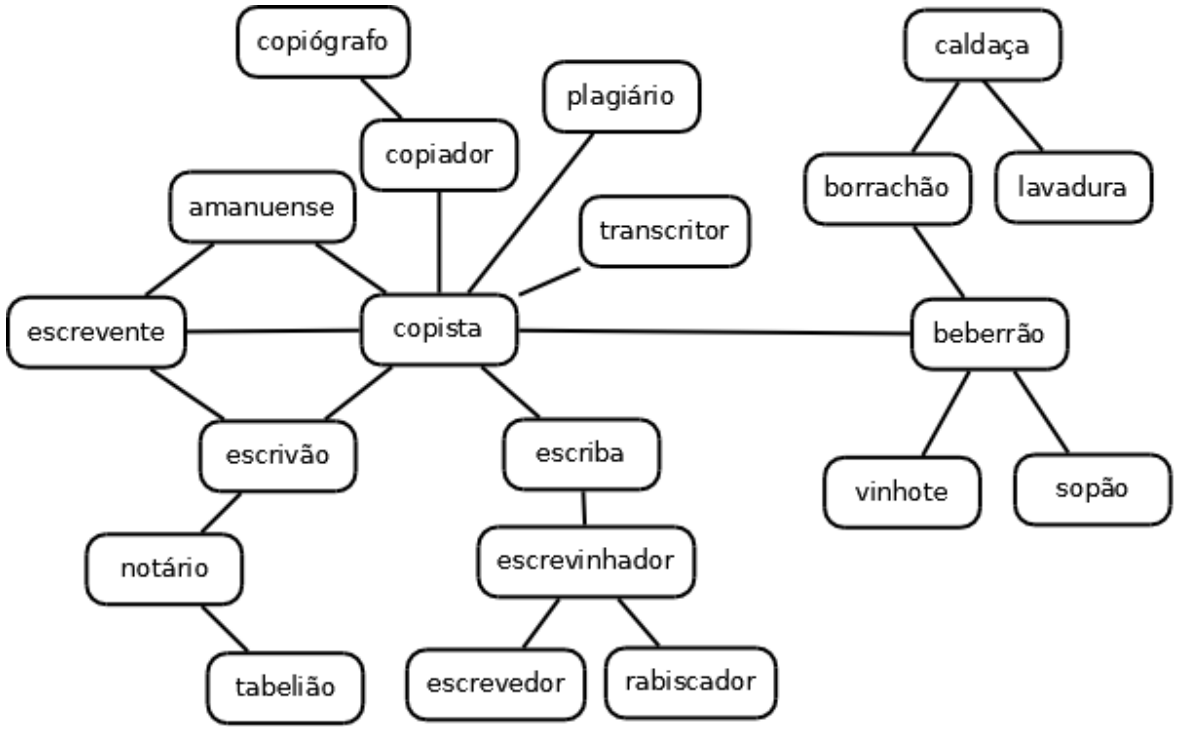
- From dictionaries:
  - `tenreiro, n -- terneiro, novilho ou bezerro.`
    → *terneiro* SYNONYM_OF *tenreiro*
    → *novilho* SYNONYM_OF *tenreiro*
    → *bezerro* SYNONYM_OF *tenreiro*
  - `ébola, n -- virose que provoca febres e hemorragias`
    → *ébola* CAUSATION_OF *febres*
    → *ébola* CAUSATION_OF *hemorragias*

Examples

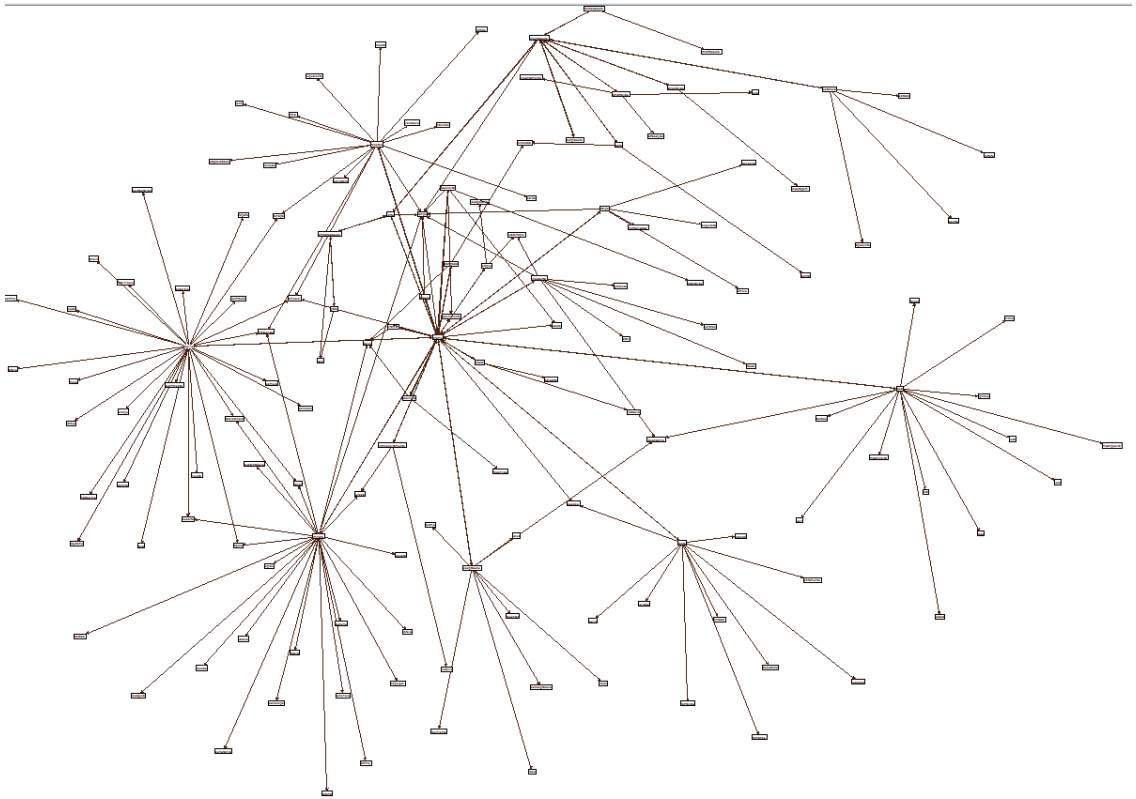
- From dictionaries:
  - `tenreiro, n -- terneiro, novilho ou bezerro.`
    → *terneiro* SYNONYM_OF *tenreiro*
    → *novilho* SYNONYM_OF *tenreiro*
    → *bezerro* SYNONYM_OF *tenreiro*
  - `ébola, n -- virose que provoca febres e hemorragias`
    → *ébola* CAUSATION_OF *febres*
    → *ébola* CAUSATION_OF *hemorragias*
- From textual corpora:
  - `O automobilismo (também conhecido como corridas de automóveis ou desporto motorizado) é um desporto...`
    → *automobilismo* SYNONYM_OF *corridas_de_automóveis*
    → *automobilismo* SYNONYM_OF *desporto_motorizado*
    → *desporto* HYPERNYM_OF *automobilismo*


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

Synonymy lexical network – example

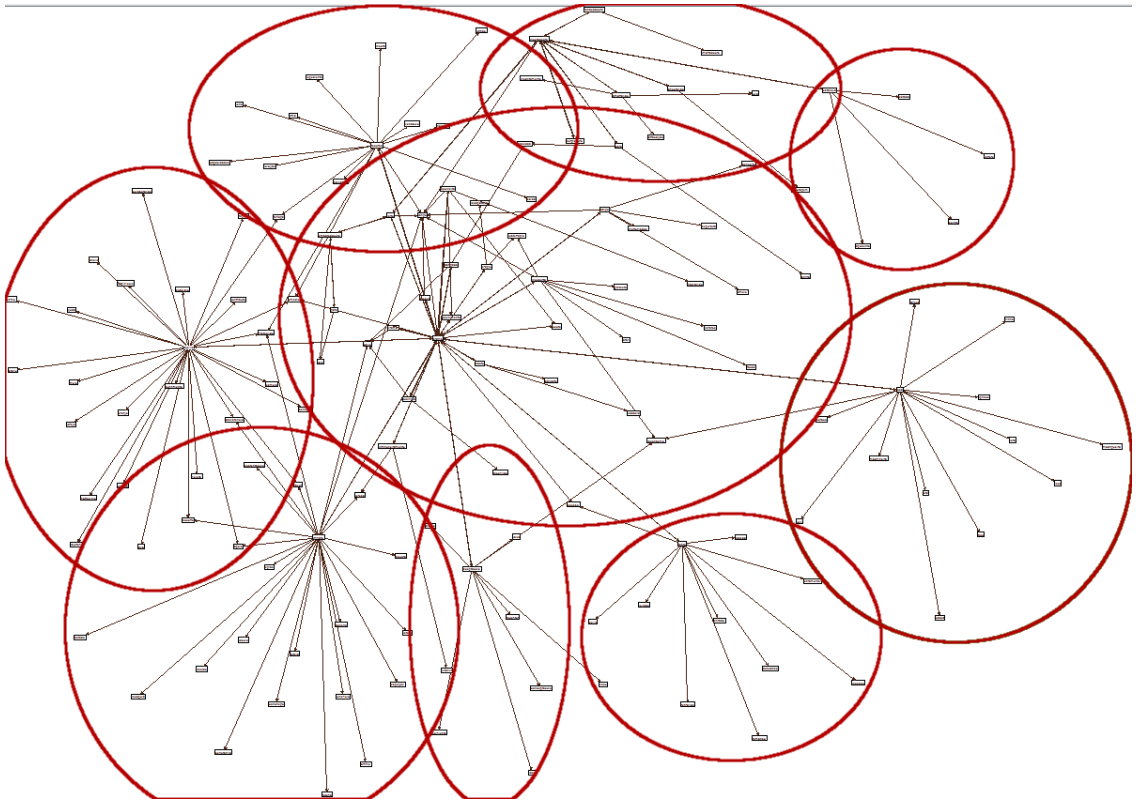


Synonymy networks tend to have a clustered structure



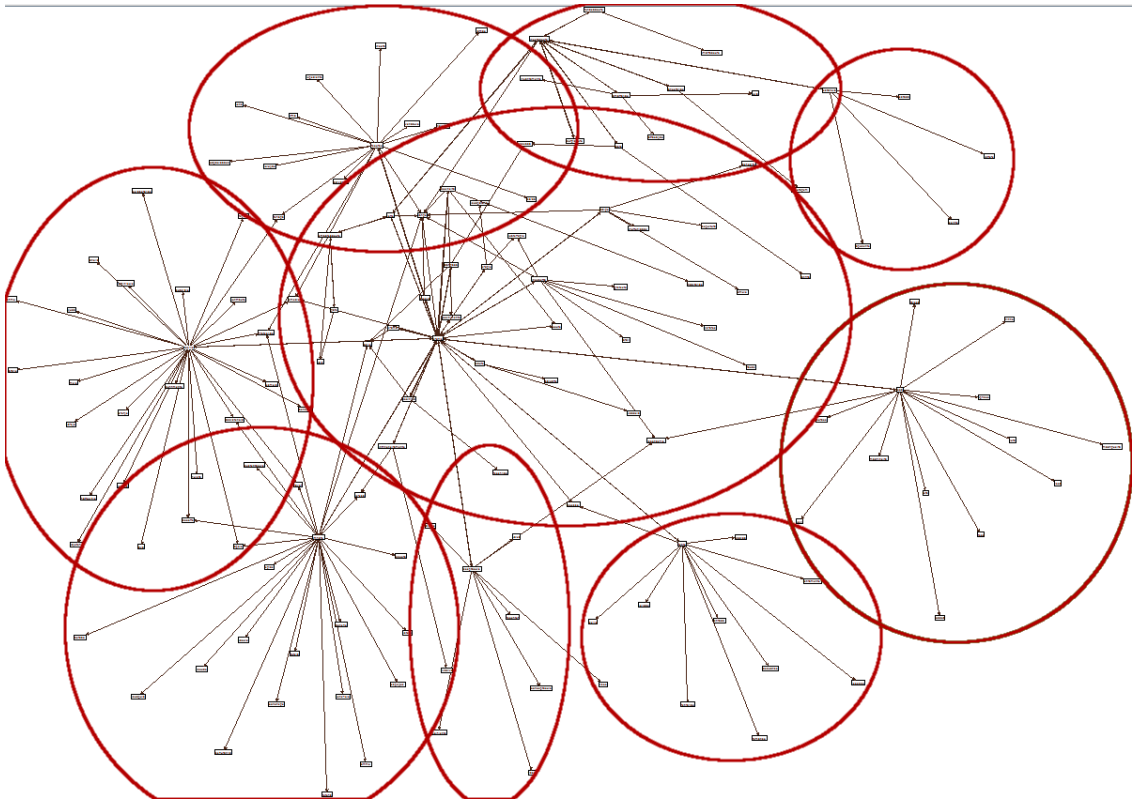
- Goal: Identify synsets taking advantage of clusters

Synonymy networks tend to have a clustered structure



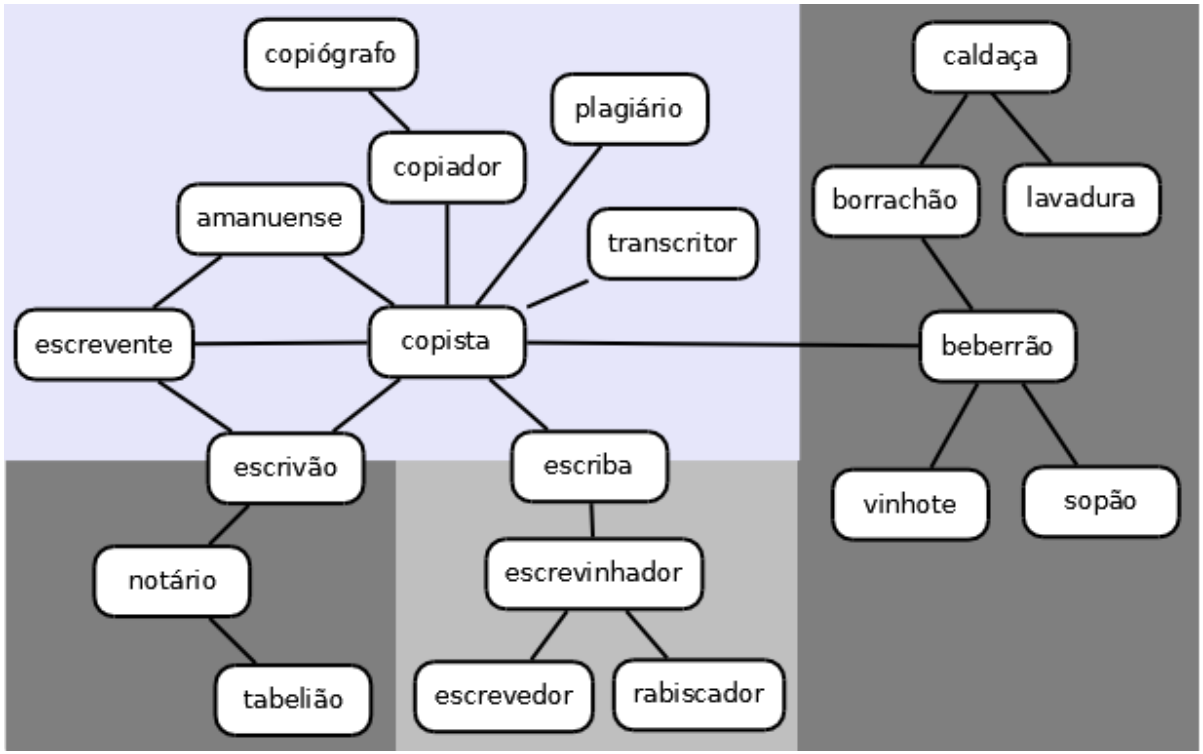
- Goal: Identify synsets taking advantage of clusters
- Approach: Clustering algorithm over the synonymy lexical network (see poster [Oliveira and Gomes, 2010])

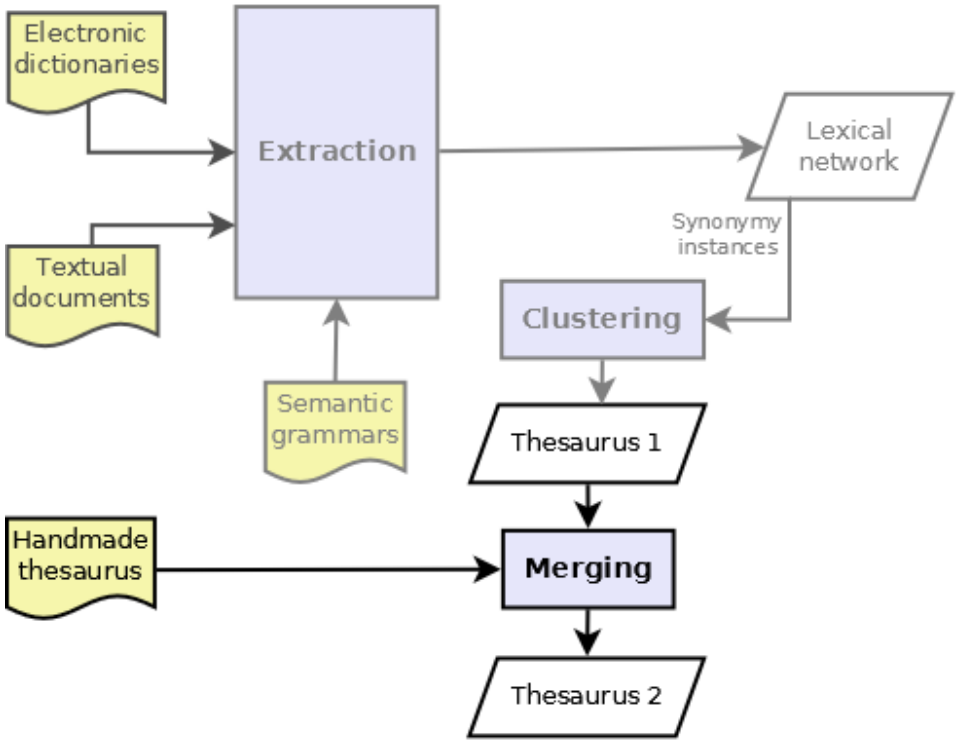
Synonymy networks tend to have a clustered structure



- Goal: Identify synsets taking advantage of clusters
- Approach: Clustering algorithm over the synonymy lexical network (see poster [Oliveira and Gomes, 2010])
- Keep ambiguity: clusters might be overlapping!

Clustering – example




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

Merging synsets from different thesauri

For each synset $T_i \in T$, select $B_j \in B$ with higher $c = T_i \cap B_j / T_i \cup B_j$[2]

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

[2] Jaccard coefficient

Merging synsets from different thesauri

For each synset $T_i \in T$, select $B_j \in B$ with higher $c = T_i \cap B_j / T_i \cup B_j$[2]

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

[2] Jaccard coefficient

Merging synsets from different thesauri

For each synset $T_i \in T$, select $B_j \in B$ with higher $c = T_i \cap B_j / T_i \cup B_j$[2]

- $B_1 = (diva, beldade, beleza, deidade, deusa, divindade)$
- $B_2 = (divindade, deidade, deus, nume)$
- $T_1 = (divindade, diva, deusa)$
  - $c(T_1, B_1) = \frac{1}{3}$
  - $c(T_1, B_2) = \frac{1}{6}$

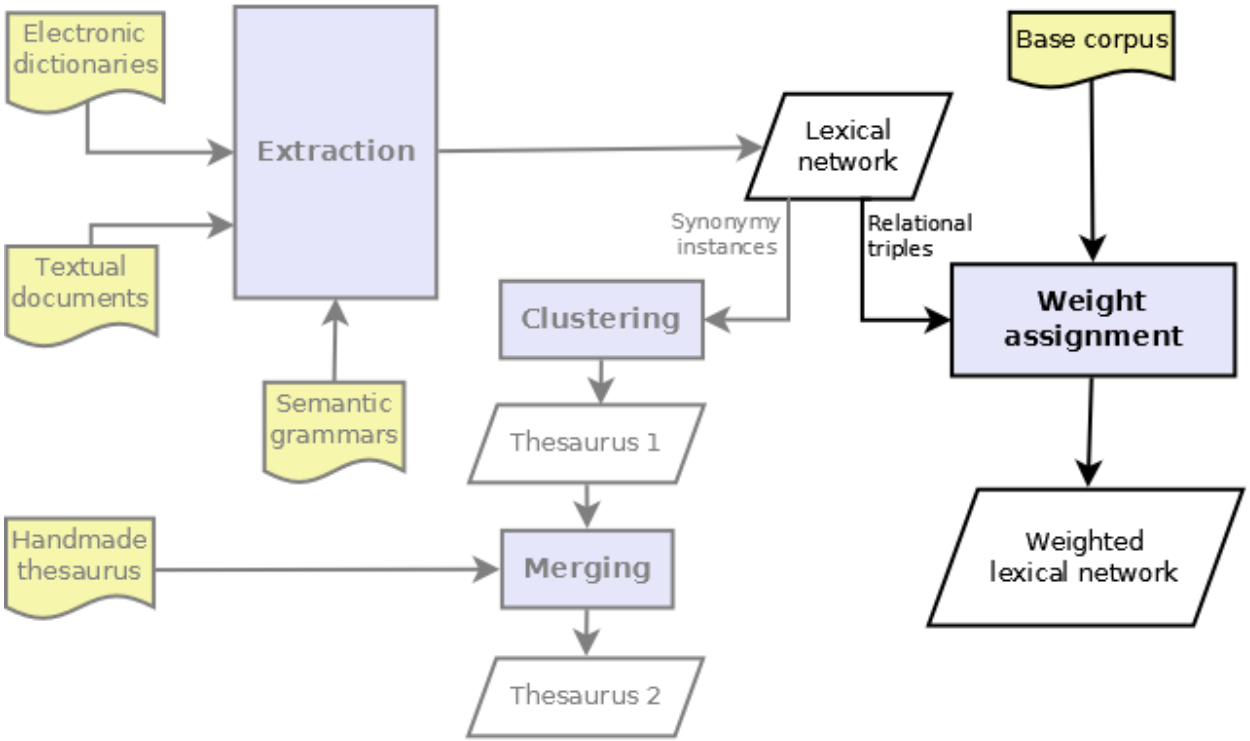
[2] Jaccard coefficient

Merging synsets from different thesauri

For each synset $T_i \in T$, select $B_j \in B$ with higher $c = T_i \cap B_j / T_i \cup B_j$[2]

- $B_1 = (diva, beldade, beleza, deidade, deusa, divindade)$
- $B_2 = (divindade, deidade, deus, nume)$
- $T_1 = (divindade, diva, deusa)$
  - $c(T_1, B_1) = \frac{1}{3}$
  - $c(T_1, B_2) = \frac{1}{6}$
- $N = B_1 \cup T_1 = (diva, beldade, beleza, deidade, deusa, divindade)$

[2] Jaccard coefficient


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

Weighting triples

- Frequency of extraction

Weighting triples

- Frequency of extraction
- Corpus distributional similarity metrics (e.g. LSA [Deerwester et al., 1990], PMI [Turney, 2001])

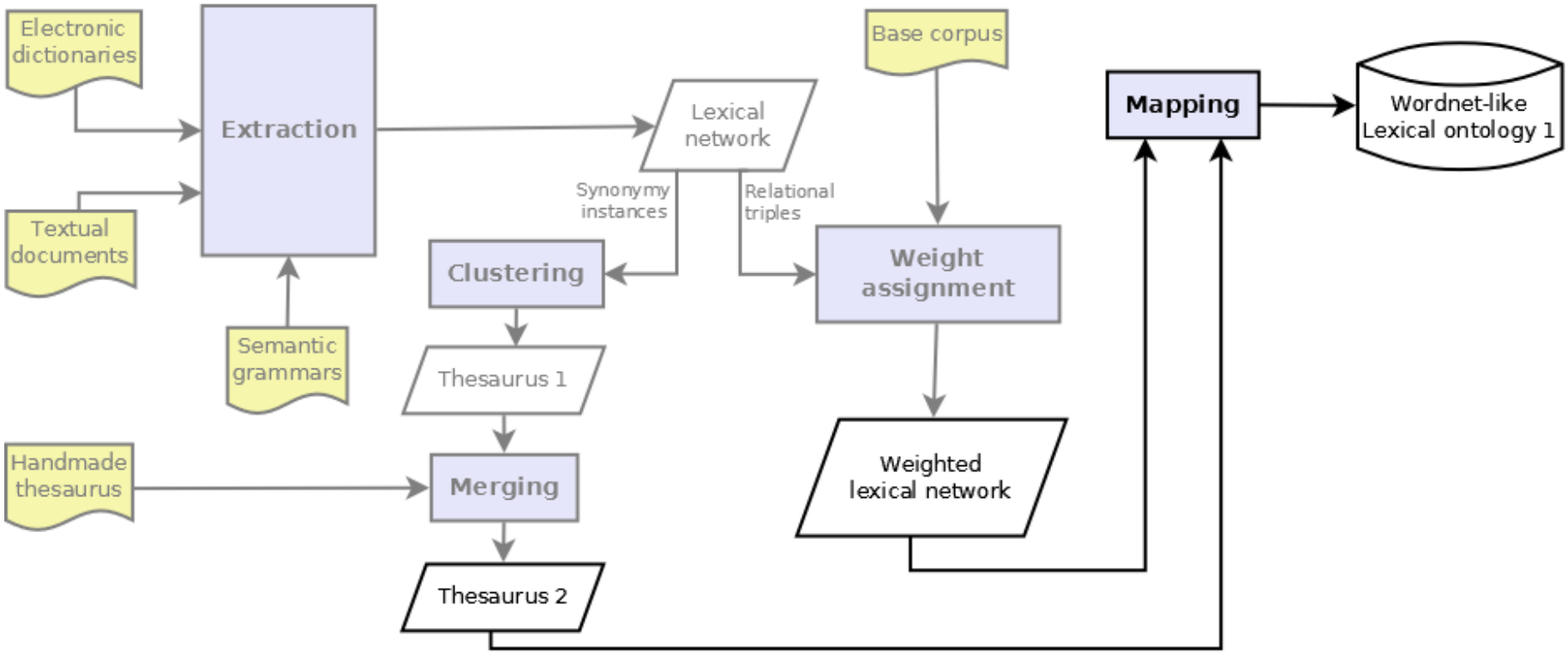
Weighting triples

- Frequency of extraction
- Corpus distributional similarity metrics (e.g. LSA [Deerwester et al., 1990], PMI [Turney, 2001])
- Web distributional similarity metrics (e.g. WebJaccard, WebOverlap [Bollegala et al., 2007])

Weighting triples

- Frequency of extraction
- Corpus distributional similarity metrics (e.g. LSA [Deerwester et al., 1990], PMI [Turney, 2001])
- Web distributional similarity metrics (e.g. WebJaccard, WebOverlap [Bollegala et al., 2007])

- Define filters based on weights


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

Mapping methods

- Input:
  - Thesaurus T, containing synsets
  - Term-based semantic network, N, where each edge has a type R

Mapping methods

- Input:
  - Thesaurus T, containing synsets
  - Term-based semantic network, N, where each edge has a type R
- Goal: map $a\ R\ b \in N$ to $A\ R\ B$, $(A, B) \in T$

Mapping methods

- Input:
  - Thesaurus T, containing synsets
  - Term-based semantic network, N, where each edge has a type R
- Goal: map $a\ R\ b \in N$ to $A\ R\ B$, $(A, B) \in T$
- Output: semantic network W, whose nodes are synsets, which relate to other synsets by means of semantic relations (wordnet)

Mapping procedures

- Baseline
  - A and B are random synsets containing a and b respectivelly

Mapping procedures

- Baseline
  - A and B are random synsets containing a and b respectivelly
- Related proportion [Gonçalo Oliveira and Gomes, 2010]

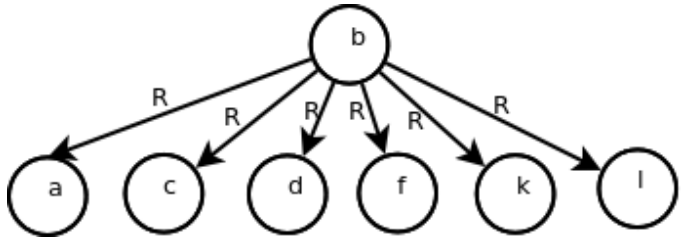
Mapping procedures

- Baseline
  - A and B are random synsets containing a and b respectivelly
- Related proportion [Gonçalo Oliveira and Gomes, 2010]
- Cosine similarity

Related proportion

Assignment of *a* (in *a R b*) to *A*:

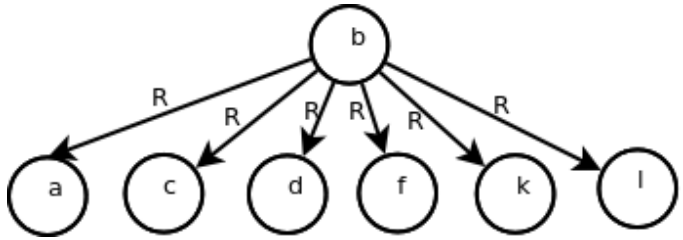
1. Fix *b*



Related proportion

Assignment of a (in $a\ R\ b$) to A:

1. Fix b

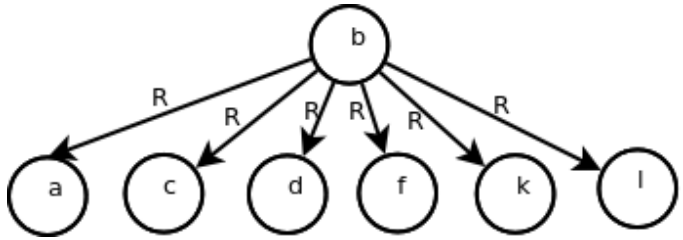


2. $S_a \subset T : S_{ai} \in S_a, a \in S_{ai}$

Related proportion

Assignment of a (in a R b) to A:

1. Fix b

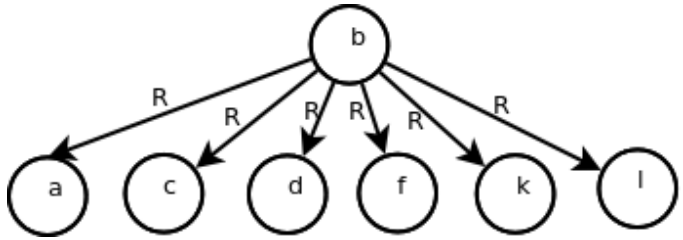


2. $S_a \subset T : S_{ai} \in S_a, a \in S_{ai}$
   - a is not in T ? create synset $A = (a)$, $a \rightarrow A$

Related proportion

Assignment of a (in a R b) to A:

1. Fix b

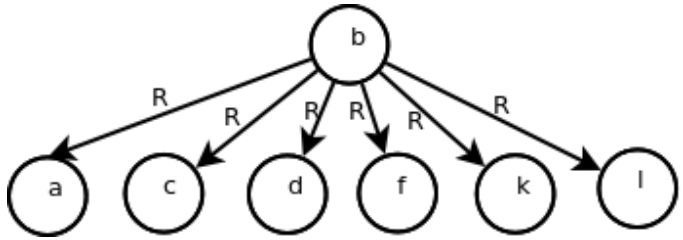


2. $S_a \subset T : S_{ai} \in S_a, a \in S_{ai}$
   - a is not in T ? create synset $A = (a)$, $a \rightarrow A$
3. For each $S_{ai} \in S_a$,

Related proportion

Assignment of a (in $a\ R\ b$) to A:

1. Fix b

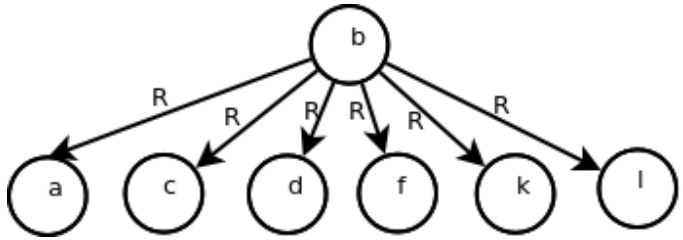


2. $S_a \subset T : S_{ai} \in S_a, a \in S_{ai}$
   - a is not in T ? create synset $A = (a)$, $a \rightarrow A$
3. For each $S_{ai} \in S_a$,
   - $p_{ai} = \frac{n_{ai}}{|S_{ai}|}$, n_{ai} = number of terms $t_j \in S_{ai} : (t_j\ R\ b)$

Related proportion

Assignment of a (in $a\ R\ b$) to A:

1. Fix b

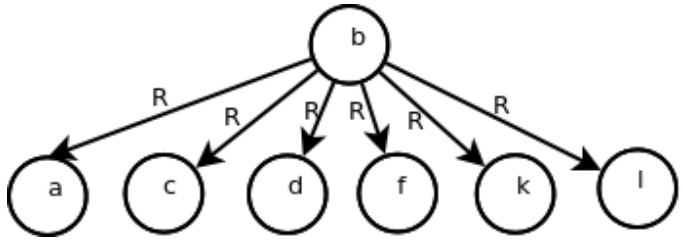


2. $S_a \subset T : S_{ai} \in S_a, a \in S_{ai}$
   - a is not in T ? create synset $A = (a)$, $a \rightarrow A$
3. For each $S_{ai} \in S_a$,
   - $p_{ai} = \frac{n_{ai}}{|S_{ai}|}$, n_{ai} = number of terms $t_j \in S_{ai} : (t_j\ R\ b)$
     - ★ $S_{a1} = (\mathbf{a}, \mathbf{c}, \mathbf{d}, e)$, $p_{a1} = \frac{3}{4}$
     - ★ $S_{a2} = (\mathbf{a}, \mathbf{f}, g)$, $p_{a2} = \frac{2}{3}$
     - ★ $S_{a3} = (\mathbf{a}, h, i, j)$, $p_{a3} = \frac{1}{4}$

Related proportion

Assignment of a (in $a\ R\ b$) to A:

1. Fix b



2. $S_a \subset T : S_{ai} \in S_a, a \in S_{ai}$
   - a is not in T ? create synset $A = (a)$, $a \rightarrow A$
3. For each $S_{ai} \in S_a$,
   - $p_{ai} = \frac{n_{ai}}{|S_{ai}|}$, n_{ai} = number of terms $t_j \in S_{ai} : (t_j\ R\ b)$
     - ★ $S_{a1} = (\mathbf{a}, \mathbf{c}, \mathbf{d}, e)$, $p_{a1} = \frac{3}{4}$
     - ★ $S_{a2} = (\mathbf{a}, \mathbf{f}, g)$, $p_{a2} = \frac{2}{3}$
     - ★ $S_{a3} = (\mathbf{a}, h, i, j)$, $p_{a3} = \frac{1}{4}$
   - $a \rightarrow S_{a1}$

Cosine similarity

1. M = *term-term* matrix based on the adjacencies of the lexical network

Cosine similarity

1. M = *term-term* matrix based on the adjacencies of the lexical network
2. Collect all the synsets with a, $S_a \subset T$, and all synsets with b, $S_b \subset T$

Cosine similarity

1. M = *term-term* matrix based on the adjacencies of the lexical network
2. Collect all the synsets with a, $S_a \subset T$, and all synsets with b, $S_b \subset T$
3. For each $A \in S_a$ and $B \in S_b$, with terms $A_i \in A$ and $B_j \in B$:

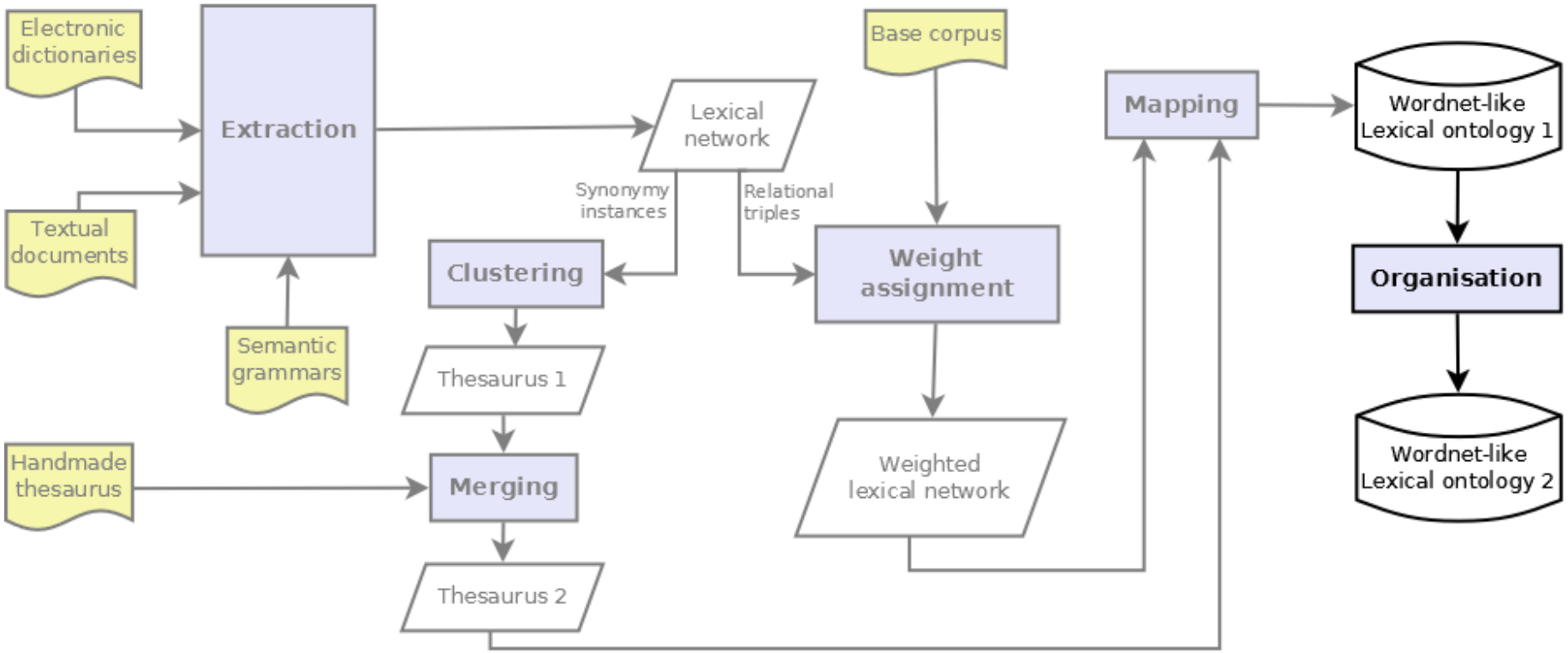
$$sim(A, B) = \frac{\sum_{i=1}^{|A|} \sum_{j=1}^{|B|} \cos(A_i, B_j)}{|A||B|}$$

Cosine similarity

1. M = *term-term* matrix based on the adjacencies of the lexical network
2. Collect all the synsets with a, $S_a \subset T$, and all synsets with b, $S_b \subset T$
3. For each $A \in S_a$ and $B \in S_b$, with terms $A_i \in A$ and $B_j \in B$:

$$sim(A, B) = \frac{\sum_{i=1}^{|A|} \sum_{j=1}^{|B|} \cos(A_i, B_j)}{|A||B|}$$

4. Select the pair of synsets with the highest similarity


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

Knowledge organisation

- Transitivity
  - if R is transitive (e.g. SYNONYMY, HYPERNYMY, ...):
    $(A\ R\ B) \wedge (B\ R\ C) \rightarrow (A\ R\ C)$

Knowledge organisation

- Transitivity
  - if R is transitive (e.g. SYNONYMY, HYPERNYMY, ...):
    $(A\ R\ B) \wedge (B\ R\ C) \rightarrow (A\ R\ C)$
- Inheritance
  - if R is not a HYPERNYMY or HYPONYMY relation:
    $(A\ \text{HYPERNYM\_OF}\ B) \wedge (A\ R\ C) \rightarrow (B\ R\ C)$

Triples extracted from dictionaries

- Dicionário da Língua Portguesa (PAPEL 2.0)
- Dicionário Aberto (DA)

Relation	Arguments	PAPEL 2.0	DA	Examples
Synonymy	noun,noun	37,452	20,910	*auxílio, contributo*
	verb,verb	21,465	8,715	*tributar, colectar*
	adj,adj	19,073	7,353	*flexível, moldável*
	adv,adv	1,171	605	*após, seguidamente*
Hypernymy	noun,noun	62,591	59,887	*planta, salva*
Part-of	noun,noun	2,805	1,795	*cauda, cometa*
	noun,adj	3.721	4,902	*tampa, coberto*
Member-of	noun,noun	5.929	1,564	*ervilha, Leguminosas*
	adj,noun	883	59	*celular, célula*
Causation	noun,noun	1.013	264	*fricção, assadura*
	adj,noun	498	166	*reactivo, reacção*
	verb,noun	6,399	5,714	*limpar, purgação*
Purpose	noun,noun	2,886	1,760	*defesa, armadura*
	verb,noun	5,192	3,383	*fazer_rir, comédia*
	verb,adj	260	186	*corrigir, correccional*

Table: Examples of triples

Relations extracted from Wikipedia abstracts

Relation	Quantity	Example	Sample	Correct	Agreement
Synonymy	11,862	*estupro,violação*	286	86,1%	91,2%
Hypernymy	29,563	*estilo_de_música,folk*	322	59,1% [3]	93,1%
Part-of	1,287	*jejuno,intestino*	268	52,6%	78,4%
Causation	520	*parasita,doença*	244	49,6%	79,5%
Purpose	743	*construção,terracota*	264	57,0%	82,2%

Table: Examples and validation of relations

[3] In 30%, grammars/tagger could not identify complete scientific names as in:
O Iriatherina werneri é uma espécie de peixe de aquário
→ *peixe_de_aquário* HYPERNYM_OF *werneri*

Thesaurus

- TeP[4] thesaurus
- OpenThesaurus.PT (OT)[5]
- Clustered PAPEL (CLIP)
- TeP merged with OT, merged with CLIP (TOP)

		TeP	OT	CLIP	TOP
Words	**Quantity**	17,158	5,819	23,741	30,554
	Ambiguous	5,867	442	12,196	13,294
	Most ambiguous	20	4	47	21
Synsets	**Quantity**	8,254	1,872	7,468	9,960
	Avg. size	3.51	3.37	12.57	6.6
	Biggest	21	14	103	277

Table: (Noun) thesauruses in numbers.

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

Manual validation

	Sample	Correct	Incorrect	N/A	Agreement
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%

Table: Results of manual synset validation.

- CLIP' and TOP' only consider synsets with 10 or less words.
  - The quality is higher for smaller synsets.

Resulting WordNet – related proportion

		Hypernym_of	Part_of	Member_of
Term-based triples		62,591	2,805	5,929
1st	**Mapped**	27,750	1,460	3,962
	Same synset	233	5	12
	Already present	3,970	40	167
Semi-mapped triples		7,952	262	357
2nd	**Mapped**	88	1	0
	Could be inferred	50	0	0
	Already present	13	0	0
Synset-based triples		23,572	1,416	3,783

Table: Results of triples mapping

Automatic validation

For each triple, *A R B*

1. Compile a set of textual patterns denoting *R*, e.g.:
   - (hypo) é `um|uma (tipo|forma|variedade|...)*` de (hyper)
   - (whole/group) é `um (grupo|conjunto|...)` de (part/member)

Automatic validation

For each triple, *A R B*

1. Compile a set of textual patterns denoting *R*, e.g.:
   - (hypo) é `um|uma (tipo|forma|variedade|...)*` de (hyper)
   - (whole/group) é `um (grupo|conjunto|...)` de (part/member)
2. Score the triple with the help of Google:

$$score = \frac{\sum_{i=1}^{|A|}\sum_{j=1}^{|B|} found(A_i, B_j, R)}{|A| * |B|}$$

Automatic validation

For each triple, *A R B*

1. Compile a set of textual patterns denoting *R*, e.g.:
   - (hypo) é `um|uma (tipo|forma|variedade|...)*` de (hyper)
   - (whole/group) é `um (grupo|conjunto|...)` de (part/member)
2. Score the triple with the help of Google:

$$score = \frac{\sum_{i=1}^{|A|}\sum_{j=1}^{|B|} found(A_i, B_j, R)}{|A| * |B|}$$

Relation	Sample size	Validation
Hypernymy_of	419 synsets	44,1%
Member_of	379 synsets	24,3%
Part_of	290 synsets	24,8%

Table: Automatic validation of triples

Concluding remarks

- Answer to the growing demand on semantically aware applications

Concluding remarks

- Answer to the growing demand on semantically aware applications
- Lack of public domain lexico-semantic resources for Portuguese

Concluding remarks

- Answer to the growing demand on semantically aware applications
- Lack of public domain lexico-semantic resources for Portuguese
- Export resources to different data formats

Concluding remarks

- Answer to the growing demand on semantically aware applications
- Lack of public domain lexico-semantic resources for Portuguese
- Export resources to different data formats

- WSD without a context:
  - Clustering for establishing synsets

Concluding remarks

- Answer to the growing demand on semantically aware applications
- Lack of public domain lexico-semantic resources for Portuguese
- Export resources to different data formats

- WSD without a context:
  - Clustering for establishing synsets
  - Disambiguation of terms based on the extracted knowledge

Concluding remarks

- Answer to the growing demand on semantically aware applications
- Lack of public domain lexico-semantic resources for Portuguese
- Export resources to different data formats

- WSD without a context:
  - Clustering for establishing synsets
  - Disambiguation of terms based on the extracted knowledge
  - Further organisation

Concluding remarks

- Answer to the growing demand on semantically aware applications
- Lack of public domain lexico-semantic resources for Portuguese
- Export resources to different data formats

- WSD without a context:
  - Clustering for establishing synsets
  - Disambiguation of terms based on the extracted knowledge
  - Further organisation
- Check `http://ontopt.dei.uc.pt` for updates and available resources

References

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.

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

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*.

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

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.

Thank you!