Extracting Lexical-Semantic Knowledge from the Portuguese Wiktionary

Leticia Antón Pérez, Hugo Gonçalo Oliveira[1], Paulo Gomes

leticiaap86@gmail.com, {hroliv,pgomes}@dei.uc.pt

Cognitive & Media Systems Group
CISUC, University of Coimbra

11 de Outubro de 2011

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

Introduction

- Online collaborative resources
  - Maintained by a community/volunteers
  - Usually public domain

[2] http://wikipedia.org
[3] http://wiktionary.org

Introduction

- Online collaborative resources
  - Maintained by a community/volunteers
  - Usually public domain
  - Huge growth potential

[2] http://wikipedia.org
[3] http://wiktionary.org

Introduction

- Online collaborative resources
  - Maintained by a community/volunteers
  - Usually public domain
  - Huge growth potential
- Wikipedia[2] – free online collaborative enclyclopedia
  - Important source of information on the world (persons, places, events)
  - Increasing popularity in text mining, information extraction and in the creation of knowledge bases [Medelyan et al., 2009]

[2] http://wikipedia.org
[3] http://wiktionary.org

Introduction

- Online collaborative resources
  - Maintained by a community/volunteers
  - Usually public domain
  - Huge growth potential
- Wikipedia[2] – free online collaborative enclyclopedia
  - Important source of information on the world (persons, places, events)
  - Increasing popularity in text mining, information extraction and in the creation of knowledge bases [Medelyan et al., 2009]
- Wiktionary[3] – free online collaborative dictionary
  - Source of lexical information
  - Less popular...

[2] http://wikipedia.org
[3] http://wiktionary.org

Lexical-semantic knowledge bases

- Structured on **words** and **meanings**
- Essential for developing NLP tools for a language

Lexical-semantic knowledge bases

- Structured on **words** and **meanings**
- Essential for developing NLP tools for a language
  - Example: Princeton WordNet [Fellbaum, 1998]

Lexical-semantic knowledge bases

- Structured on **words** and **meanings**
- Essential for developing NLP tools for a language
  - Example: Princeton WordNet [Fellbaum, 1998]
  - Created manually
  - Current on version 3.1

Lexical-semantic knowledge bases

- Structured on **words** and **meanings**
- Essential for developing NLP tools for a language
  - Example: Princeton WordNet [Fellbaum, 1998]
  - Created manually
  - Current on version 3.1
- Automatic extraction of lexical knowledge from dictionaries
  - Extraction of hierarchies (eg. [Chodorow et al., 1985])

Lexical-semantic knowledge bases

- Structured on **words** and **meanings**
- Essential for developing NLP tools for a language
  - Example: Princeton WordNet [Fellbaum, 1998]
  - Created manually
  - Current on version 3.1
- Automatic extraction of lexical knowledge from dictionaries
  - Extraction of hierarchies (eg. [Chodorow et al., 1985])
  - Creation of knowledge bases (e.g MindNet [Richardson et al., 1998])

Lexical-semantic knowledge bases

- Structured on **words** and **meanings**
- Essential for developing NLP tools for a language
  - Example: Princeton WordNet [Fellbaum, 1998]
  - Created manually
  - Current on version 3.1
- Automatic extraction of lexical knowledge from dictionaries
  - Extraction of hierarchies (eg. [Chodorow et al., 1985])
  - Creation of knowledge bases (e.g MindNet [Richardson et al., 1998])
  - Also for Portuguese (eg. PAPEL [Gonçalo Oliveira et al., 2010])

Goal

- **Extract lexical-semantic knowledge from the Wiktionary.PT**

Goal

- **Extract lexical-semantic knowledge from the Wiktionary.PT**
  1. Process the Wiktionary XML files

Goal

- **Extract lexical-semantic knowledge from the Wiktionary.PT**
  1. Process the Wiktionary XML files
  2. Analyse the vocabulary of the definitions

Goal

- **Extract lexical-semantic knowledge from the Wiktionary.PT**
  1. Process the Wiktionary XML files
  2. Analyse the vocabulary of the definitions
  3. Select a list of semantic relations to extract (eg. synonymy, part-of, ...)

Goal

- **Extract lexical-semantic knowledge from the Wiktionary.PT**
  1. Process the Wiktionary XML files
  2. Analyse the vocabulary of the definitions
  3. Select a list of semantic relations to extract (eg. synonymy, part-of, ...)
  4. Develop extraction grammars

Goal

- **Extract lexical-semantic knowledge from the Wiktionary.PT**
  1. Process the Wiktionary XML files
  2. Analyse the vocabulary of the definitions
  3. Select a list of semantic relations to extract (eg. synonymy, part-of, ...)
  4. Develop extraction grammars
  5. Extract instances of semantic relations between lemmas
     (eg. *animal* hypernym-of *cão*)

Goal

- **Extract lexical-semantic knowledge from the Wiktionary.PT**
  1. Process the Wiktionary XML files
  2. Analyse the vocabulary of the definitions
  3. Select a list of semantic relations to extract (eg. synonymy, part-of, ...)
  4. Develop extraction grammars
  5. Extract instances of semantic relations between lemmas
     (eg. *animal* hypernym-of *cão*)
  6. Validate the results

Goal

- **Extract lexical-semantic knowledge from the Wiktionary.PT**
  1. Process the Wiktionary XML files
  2. Analyse the vocabulary of the definitions
  3. Select a list of semantic relations to extract (eg. synonymy, part-of, ...)
  4. Develop extraction grammars
  5. Extract instances of semantic relations between lemmas (eg. *animal* hypernym-of *cão*)
  6. Validate the results
- Useful for..
  - NLP tasks (eg. word sense disambiguation, question-answering, writing aids ...)

Goal

- **Extract lexical-semantic knowledge from the Wiktionary.PT**
  1. Process the Wiktionary XML files
  2. Analyse the vocabulary of the definitions
  3. Select a list of semantic relations to extract (eg. synonymy, part-of, ...)
  4. Develop extraction grammars
  5. Extract instances of semantic relations between lemmas (eg. *animal* hypernym-of *cão*)
  6. Validate the results
- Useful for..
  - NLP tasks (eg. word sense disambiguation, question-answering, writing aids ...)
  - Augment existing knowledge bases (eg. PAPEL)

Wiktionary files

XML+wiki

```
<page>
<title>computador</title>
...
<text xml:space="preserve≫wikipedia
=Português=
==Adjetivo==
flex.pt|ms=computador|mp=computadores|fs=computadora|fp
=computadoras
oxítona|com|pu|ta|dor
# que [[computar|computa]]
==Substantivo==
flex.pt|ms=computador|mp=computadores
oxítona|com|pu|ta|dor
# o que [[fazer|faz]] [[cômputo]]s ([[cálculo]]s);
o que [[computar|computa]]
# [[máquina]] [[capaz]] de [[fazer]] [[cálculo]]s
===Sinônimos===
* De '''1''':  [[calculista]]
* De '''2''':  [[calculadora]]
===Tradução===
tradini|De 3 (aparelho eletrônico capaz de calcular)
* trad|af|rekenaar
...
</page>
```

Wiktionary files

XML+wiki

```
<page>
<title>computador</title>
...
<text xml:space="preserve≫wikipedia
=Português=
==Adjetivo==
flex.pt|ms=computador|mp=computadores|fs=computadora|fp
=computadoras
oxítona|com|pu|ta|dor
# que [[computar|computa]]
==Substantivo==
flex.pt|ms=computador|mp=computadores
oxítona|com|pu|ta|dor
# o que [[fazer|faz]] [[cômputo]]s ([[cálculo]]s);
o que [[computar|computa]]
# [[máquina]] [[capaz]] de [[fazer]] [[cálculo]]s
===Sinônimos===
* De '''1''':  [[calculista]]
* De '''2''':  [[calculadora]]
===Tradução===
tradini|De 3 (aparelho eletrônico capaz de calcular)
* trad|af|rekenaar
...
</page>
```

- Over 170,000 entries ($\approx$ 110,000 refer to Portuguese words)

Definitions collected

Definitions file

computador adj que computa
computador nome o que faz cômputos (cálculos); o que computa
computador nome máquina capaz de fazer cálculos
computador adj calculista
computador nome calculadora

Definitions collected

Definitions file

computador adj que computa
computador nome o que faz cômputos (cálculos); o que computa
computador nome máquina capaz de fazer cálculos
computador adj calculista
computador nome calculadora

POS	Definitions[4]	Examples
Nouns	41,836	*homem: um tipo de primata bípede e bímano da espécie Homo Sapiens* (man: a kind of primate bipedal and bimane of the Homo Sapiens species)
Verbs	8, 703	*avermelhar: fazer ou tornar vermelho* (redden: to make or to become red)
Adjectives	14,987	*bípede: que tem dois membros* (biped: with two members)
Adverbs	909	*moralmente: de maneira moral* (morally: in a moral way)

[4] Including synonymy lists; excluding inflected verbs and closed category words

Vocabulary analysis

N-gram	Frequency	Part-of-speech	Semantic relation
o mesmo que (the same as)	756	Noun	Synonymy
ato ou efeito de (act or effect of)	435	Noun	Causation
conjunto de (set of)	248	Noun	Member-of
pessoa que (person who)	237	Noun	Hypernymy
espécie de (species of)	175	Noun	Hypernymy
o mesmo que (the same as)	72	Verb	Synonymy
relativo à/ao (relative to)	619	Adjective	Property
que se (that)	399	Adjective	Property
que tem (that has)	382	Adjective	Part-of/Property
diz-se de (it is said about)	245	Adjective	Property
habitante ou natural de (inhabitant or natural of)	143	Adjective	Place/Origin
de modo (in a way)	82	Adverb	Manner
de maneira (in a manner)	26	Adverb	Manner

Covered by the grammars of PAPEL (www.linguateca.pt/PAPEL/)

Extraction procedure

1. Manual creation of extraction grammars

```
...
ROOT ::= ...  <&> usado <&> para <&> PURPOSE_OF
ROOT ::= parte <&> DE_DO_DA <&> HAS_PART
ROOT ::= ...  <&> que <&> contém <&> DET <&> PART_OF

...
```

Extraction procedure

1. Manual creation of extraction grammars

```
...
ROOT ::= ...  <&> usado <&> para <&> PURPOSE_OF
ROOT ::= parte <&> DE_DO_DA <&> HAS_PART
ROOT ::= ...  <&> que <&> contém <&> DET <&> PART_OF

...
```

2. Automatic extraction of semantic relation instances

```
candeia  nome  utensílio doméstico rústico usado para iluminação , com pavio abastecido a óleo
→ com purpose-of candeia
→ iluminação purpose-of candeia
espiga  nome  parte das gramíneas que contém os grãos
→ espiga part-of gramíneas
→ grãos part-of espiga

...
```

Extraction procedure

1. Manual creation of extraction grammars

```
...
ROOT ::= ...  <&> usado <&> para <&> PURPOSE_OF
ROOT ::= parte <&> DE_DO_DA <&> HAS_PART
ROOT ::= ...  <&> que <&> contém <&> DET <&> PART_OF

...
```

2. Automatic extraction of semantic relation instances

```
candeia  nome  utensílio doméstico rústico usado para iluminação , com pavio abastecido a óleo
→ com purpose-of candeia
→ iluminação purpose-of candeia
espiga  nome  parte das gramíneas que contém os grãos
→ espiga part-of gramíneas
→ grãos part-of espiga

...
```

3. Automatic adjustment of relations and argument lemmatisation

```
candeia  nome  utensílio#n doméstico#adj rústico#adj usado#v-pcp para#prp iluminação#n ,#punc com#prp
pavio#n abastecido#v-pcp a#prp óleo#n
→ iluminação purpose-of candeia
espiga  nome  parte#n de#prp as#art gramíneas#n que#pron-indp contém#v-fin os#art grãos#n
→ espiga part-of gramínea
→ grão part-of espiga

...
```

Extraction results

- 55,705 relational triples are were extracted

Extraction results

- 55,705 relational triples are were extracted

Group	Name	Args.	Qnt.	Examples
Synonymy	SINONIMO_N_DE	n,n	13,647	*léxico*,*dicionário* (lexicon,dictionary)
	SINONIMO_V_DE	v,v	4,136	*esticar*,*estender* (to_extend,to_stretch)
	SINONIMO_ADJ_DE	adj,adj	6,660	*pronto*,*súbito* (prompt,sudden)
Hypernymy	HIPERONIMO_DE	n,n	18,638	*pessoa*,*guerreiro* (person,warrior)
Part-of	PARTE_DE	n,n	723	*núcleo*,*átomo* (core,atom)
	PARTE_DE_ALGO_COM_PROPRIEDADE	n,adj	559	*vício*,*vicioso* (addiction,addictive)
Member-of	MEMBRO_DE	n,n	1,166	*aluno*,*escola* (student,school)
Causation-of	ACCAO_QUE_CAUSA	v,n	1,149	*mover*,*movimento* (to_move,movement)
	CAUSADOR_DE	n,n	307	*vírus*,*doença* (virus,disease)
Producer-of	PRODUTOR_DE	n,n	316	*oliveira*,*azeitona* (olive_tree,olive)
Purpose-of	ACCAO_FINALIDADE_DE	v,n	1,485	*calcular*,*cálculo* (to_calculate,calculation)
	FINALIDADE_DE	n,n	1,355	*sustentação*,*mastro* (support,mast)
Location	LOCAL_ORIGEM_DE	n,n	747	*Índia*,*hindu* (India,hindu)

Corpus support

- Search for sentences in CETEMPúblico [Rocha and Santos, 2000] supporting the relations between nouns

Corpus support

- Search for sentences in CETEMPúblico [Rocha and Santos, 2000] supporting the relations between nouns

Relation	Sentence
língua HIPERONIMO_DE *alemão*	*As iniciativas deste gabinete passam geralmente pela promoção de conferências, exposições, workshops e aulas de* **línguas, como o inglês, alemão** *ou japonês.*
ciência HIPERONIMO_DE *grafologia*	*Mas para Alberto Vaz da Silva,* **a grafologia é uma ciência** *que, além de definir o carácter e temperamento de um indivíduo, pode ajudá-lo a libertar-se de culpas e complexos ganhos na infância.*
rua PARTE_DE *quarteirão*	*...* **quarteirão formado pelas ruas** *de Rodrigues de Freitas, dos Polacos e de Marciano Aziaga, ...*
mercúrio PARTE_DE *amálgama*	*O mercúrio é uma substância altamente tóxica e as* **amálgamas dentárias são feitas de mercúrio**.
pessoa MEMBRO_DE *comissão*	*A* **comissão é constituída por pessoas** *que ficaram marcadas pela presença de Dona Amélia: ...*
lobo MEMBRO_DE *alcateia*	*Mech e os seus colegas constataram que alguns dos cheiros contidos nas marcas de urina servem para os* **lobos de uma alcateia** *saberem por onde andou o lobo que deixou as marcas ...*
transporte FINALIDADE_DE *embarcação*	*... onde foi descoberto o resto do casco de uma* **embarcação presumivelmente utilizada no transporte** *de peças de cerâmica ...*
vírus CAUSADOR_DE *doença*	*A hepatite A transmite-se enquanto as pessoas não têm sintomas, é uma* **doença benigna, provocada por um vírus** *que causa fraqueza, incómodos, febre e vómitos ...*

Comparison with PAPEL 2.0 (1)

- Support for Wiktionary relations

Relation	Freq(args)[5] ≥ 100				Cooc(args) ≥ 1[6]			
	Total		Supported		Total		Supported	
Hypernymy	6,556	35.2%	2,074	31.6%	6,584	36.9%	2,249	34.2%
Part-of	235	32.5%	91	38.7%	238	33.1%	99	41.6%
Member-of	323	27.7%	144	44.6%	329	28.6%	149	45.3%
Causation	99	32.2%	14	14.1%	82	26.8%	12	14.6%
Purpose-of	440	32.5%	66	15.0%	445	33.0%	70	15.7%

[5] **Freq(args)**: only instances whose arguments occur more than 100 times in CETEMPúblico

[6] **Cooc(args)**: only instances whose arguments co-occur at least once in a sentence

Comparison with PAPEL 2.0 (1)

- Support for Wiktionary relations

Relation	Freq(args)[5] ≥ 100				Cooc(args) ≥ 1[6]			
	Total		Supported		Total		Supported	
Hypernymy	6,556	35.2%	2,074	31.6%	6,584	36.9%	2,249	34.2%
Part-of	235	32.5%	91	38.7%	238	33.1%	99	41.6%
Member-of	323	27.7%	144	44.6%	329	28.6%	149	45.3%
Causation	99	32.2%	14	14.1%	82	26.8%	12	14.6%
Purpose-of	440	32.5%	66	15.0%	445	33.0%	70	15.7%

- Support for PAPEL 2.0 relations

Relation	Freq(args) ≥ 100				Cooc(args) ≥ 1			
	Total		Supported		Total		Supported	
Hypernymy	17,749	28.4%	5,196	29.3%	18,511	29.6%	5,749	31.1%
Part-of	625	22.3%	183	29.3%	666	23.7%	212	31.8%
Member-of	1,035	17.5%	445	43.0%	1150	19.4%	462	40.2%
Causation	227	22.4%	35	15.4%	218	21.5%	41	18.8%
Purpose-of	839	29.1%	126	15.0%	855	29.6%	134	15.7%

[5] **Freq(args)**: only instances whose arguments occur more than 100 times in CETEMPúblico

[6] **Cooc(args)**: only instances whose arguments co-occur at least once in a sentence

Comparison with PAPEL 2.0 (2)

Semantic relation	Args	PAPEL	Wiktionary	Common	New from Wiktionary	
Hypernymy	n,n	62,591	17,837	3,442	14,395	(+23%)
Synonymy	n,n	37,452	13,556	2,949	10,607	(+28%)
	v,v	21,465	4,076	1,275	2,801	(+13%)
	adj,adj	19,073	6,629	1,740	4,889	(+26%)
	adv,adv	1,171	289	94	195	(+17%)
Part-of	n,n	2,805	718	47	671	(+24%)
	n,adj	3,721	558	146	412	(+11%)
	adj,n	17	26	0	26	(+153%)
Member-of	n,n	5,929	1,152	83	1,069	(+18%)
	n,adj	34	23	1	22	(+65%)
Causation	n,n	1,013	306	14	292	(+29%)
	v,n	6,399	1,139	645	494	(+8%)
Producer-of	n,n	898	310	14	296	(+33%)
Purpose-of	n,n	2,886	1,349	46	1,303	(+45%)
	v,n	5,192	1,479	126	1,353	(+26%)
Place-of	n,n	849	722	1	721	(+85%)
Manner-of	adv,n	1,113	156	64	92	(+8%)
Property-of	adj,v	17,543	3,239	404	2,835	(+16%)
	adj,n	6,518	1,625	242	1,383	(+21%)

Concluding remarks (1)

- The Portuguese Wiktionary
  - Valuable resource in the automatic creation/enrichment of lexical knowledge bases (eg. PAPEL).

Concluding remarks (1)

- The Portuguese Wiktionary
  - Valuable resource in the automatic creation/enrichment of lexical knowledge bases (eg. PAPEL).
  - Its vocabulary enables reusing existing grammars for extracting relations from dictionaries

Concluding remarks (1)

- The Portuguese Wiktionary
  - Valuable resource in the automatic creation/enrichment of lexical knowledge bases (eg. PAPEL).
  - Its vocabulary enables reusing existing grammars for extracting relations from dictionaries
- Extracted relations freely available through http://ontopt.dei.uc.pt

Concluding remarks (2)

- This work is part of a larger project:
  - **Onto.PT**: automatic construction of a lexical ontology for Portuguese [Gonçalo Oliveira and Gomes, 2010]

Concluding remarks (2)

- This work is part of a larger project:
  - **Onto.PT**: automatic construction of a lexical ontology for Portuguese [Gonçalo Oliveira and Gomes, 2010]
- Semantic relations are extracted from several textual resources

Concluding remarks (2)

- This work is part of a larger project:
  - **Onto.PT**: automatic construction of a lexical ontology for Portuguese [Gonçalo Oliveira and Gomes, 2010]
- Semantic relations are extracted from several textual resources
- After extraction...
  - Discovery of synsets [Gonçalo Oliveira and Gomes, 2011a] / enrichment of existing thesaurus [Gonçalo Oliveira and Gomes, 2011b]
  - Integration of relations [Gonçalo Oliveira and Gomes, 2011c]

Concluding remarks (2)

- This work is part of a larger project:
  - **Onto.PT**: automatic construction of a lexical ontology for Portuguese [Gonçalo Oliveira and Gomes, 2010]
- Semantic relations are extracted from several textual resources
- After extraction...
  - Discovery of synsets [Gonçalo Oliveira and Gomes, 2011a] / enrichment of existing thesaurus [Gonçalo Oliveira and Gomes, 2011b]
  - Integration of relations [Gonçalo Oliveira and Gomes, 2011c]
- Collaborative resources keep growing and so will the results obtained from their exploitation.

Thank you!

References I

[Chodorow et al., 1985] Chodorow, M. S., Byrd, R. J., and Heidorn, G. E. (1985).
Extracting semantic hierarchies from a large on-line dictionary.
In *Proceedings of the 23rd annual meeting on Association for Computational Linguistics*, pages 299–304, Morristown, NJ, USA. Association for Computational Linguistics.

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

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

[Gonçalo Oliveira and Gomes, 2011a] Gonçalo Oliveira, H. and Gomes, P. (2011a).
Automatic discovery of fuzzy synsets from dictionary definitions.
In *Proceedings 22nd International Joint Conference on Artificial Intelligence (IJCAI 2011)*, pages 1801–1806, Barcelona, Spain. AAAI Press.

[Gonçalo Oliveira and Gomes, 2011b] Gonçalo Oliveira, H. and Gomes, P. (2011b).
Automatically enriching a thesaurus with information from dictionaries.
In *Proceedings 15th Portuguese Conference on Artificial Intelligence (EPIA 2011)*, Lisbon, Portugal. APPIA.

[Gonçalo Oliveira and Gomes, 2011c] Gonçalo Oliveira, H. and Gomes, P. (2011c).
Ontologising relational triples into a portuguese thesaurus.
In *Local Proceedings 15th Portuguese Conference on Artificial Intelligence (EPIA 2011)*, Lisbon, Portugal. APPIA.

[Gonçalo Oliveira et al., 2010] 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.

References II

[Medelyan et al., 2009] Medelyan, O., Milne, D., Legg, C., and Witten, I. H. (2009).
Mining meaning from wikipedia.
Intl. Journal of Human-Computer Studies.

[Richardson et al., 1998] Richardson, S. D., Dolan, W. B., and Vanderwende, L. (1998).
Mindnet: Acquiring and structuring semantic information from text.
In *Proceedings of 17th International Conference on Computational Linguistics (COLING)*, pages 1098–1102.

[Rocha and Santos, 2000] Rocha, P. A. and Santos, D. (2000).
CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa.
In das Graças Volpe Nunes, M., editor, *V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR 2000)*, pages 131–140, São Paulo. ICMC/USP.