Ontology Learning for Portuguese

Hugo Gonçalo Oliveira*
hroliv@dei.uc.pt

NLIP Seminar Series, University of Cambridge

University of Coimbra
Faculty of Sciences and Technology
Department of Informatics Engineering

Knowledge and Intelligent Systems Laboratory
Cognitive and Media Systems Group
Centre of Informatics and Systems of the University of Coimbra

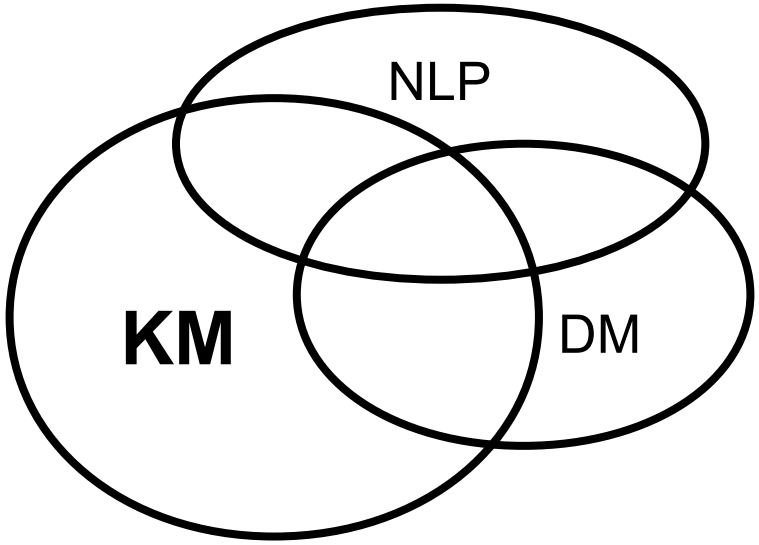
* Hugo Gonçalo Oliveira is supported by FCT scholarship, grant SFRH/BD/44955/2008.

Outline

- My research group
- PhD work: Ontology Learning for Portuguese
  - Lexical resources
  - Goals
  - Motivation
  - Approach
  - Evaluation
  - Concluding remarks

KIS Research Group

- Knowledge and Intelligent Systems
  - Knowledge Management (KM)
  - Natural Language Processing (NLP)
  - Data Mining (DM)



KIS Research Group - NLP

- Works going on
  - Onto.PT: Ontology Learning from Portuguese text
  - RAPPORT: Automatic Question & Answering for Portuguese
  - Web Page Annotation
- Former node of Linguateca
  - PAPEL: a lexical ontology created semi-automatically from a general dictionary
  - Floresta Sintá(c)tica: a treebank for Portuguese

Linguateca

- A distributed resource center for Portuguese language technology
  - http://www.linguateca.pt
- Government funded initiative to significantly raise the quality and availability of resources for the computational processing of Portuguese
- Network headed by a small group (Linguateca's Oslo node) at SINTEF ICT

Linguateca
Main goal

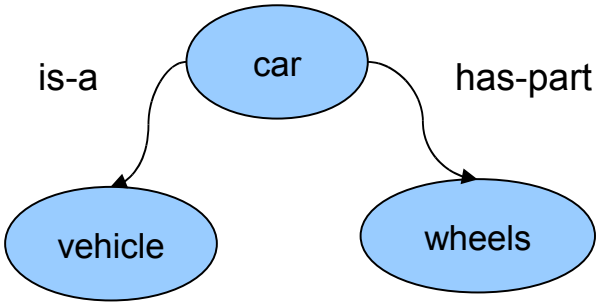
- To guarantee that:
  - Information was provided and gathered at one place on the Web
  - Resources were made public, maintained, and further developed in connection with the scientific community
  - Evaluation initiatives were launched

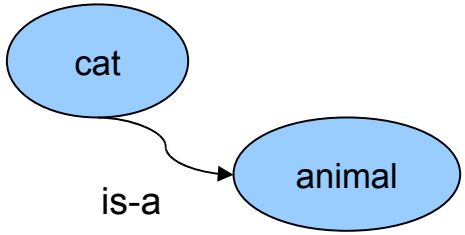
Linguateca
Achievements

- A lot of publicly available resources
- Several evaluation contests which advanced the state of the art
- Information, dissemination, gathering of relevant data and a team who answers
- The first evaluation contest for Portuguese
- The first treebank for Portuguese
- The first Web-based corpus service for Portuguese
- The first Q&A system for Portuguese
- The largest revised and annotated parallel corpus in the world
- The first national Web snapshot available

Ontology Learning for Portuguese

- *The cat has four wheels.*
  - Spelling: correct
  - Syntax: correct
  - Semantics: at least... strange!
- Natural language processing (NLP) needs access to semantic resources





Lexical resources

- Dictionaries
- Thesaurus
- Taxonomies
- Lexical knowledge bases/ontologies
  - Conceptual models of words and their meanings
  - Lexico-semantic relations
    - Synonymy: *car* syn *auto*
    - Hyponymy (is-a): *ambulance* is-a *car*
    - Meronymy (part-of): *wheel* part-of *car*
    - Cause, purpose, location...

State of the art lexical ontologies
For English

- Lexical knowledge bases for English
  - Handcrafted:
    - Princeton WordNet (Fellbaum 1998)
    - Cyc (Lenat and Guha 1989)
    - Berkeley FrameNet (Baker et al. 1998)
  - Created semi-automatically:
    - MindNet (Richardson et al. 1998)

State of the art lexical ontologies
For Portuguese

- Handcrafted:
  - WordNet.PT (Marrafa 2002)
    - Closed
  - WordNet.BR (Dias da Silva et al. 2002)
    - Not yet available
  - Tep (Maziero et al. 2008)
    - Free, but only synonyms and antonyms
  - MultiWordNet.PT
    - First version, paid license

State of the art lexical ontologies
For Portuguese

- Created semi-automatically:
  - PAPEL (Gonçalo Oliveira et al. 2008, 2009)
    - Extracted semi-automatically from a general Portuguese dictionary
    - Using handcrafted semantic grammars
    - Freely available by Linguateca (http://www.linguateca.pt/PAPEL)

PAPEL

- PAPEL 1.1: about 200,000 relational triples between Portuguese terms

Group	Name	Args.	Qnt.	Examples
Synonymy	SINONIMO_N_DE	n,n	37,259	(*auxílio*, *contributo*)
	SINONIMO_V_DE	v,v	21,534	(*tributar*, *colectar*)
	SINONIMO_ADJ_DE	adj,adj	19,073	(*flexível*, *moldável*)
	SINONIMO_ADV_DE	adv,adv	1,169	(*após*, *seguidamente*)
Hypernymy	HIPERONIMO_DE	n,n	61,477	(*planta*, *salva*)
Meronymy	PARTE_DE	n,n	9,970	(*cauda*, *cometa*)
	PARTE_DE_ALGO_COM_PROP	n,adj	3,806	(*tampa*, *coberto*)
	PROP_DE_ALGO_PARTE_DE	adj,n	900	(*celular*, *célula*)
Cause	CAUSADOR_DE	n,n	1,010	(*fricção*, *assadura*)
	CAUSADOR_DE_ALGO_COM_PROP	n,adj	17	(*paixão*, *passional*)
	PROP_DE_ALGO_CAUSADOR_DE	adj,n	498	(*reactivo*, *reacção*)
	ACCAO_QUE_CAUSA	v,n	6,399	(*limpar*, *purgação*)
	CAUSADOR_DA_ACCAO	n,v	39	(*gases*, *fumigar*)
Producer	PRODUTOR_DE	n,n	885	(*romãzeira*, *romã*)
	PRODUTOR_DE_ALGO_COM_PROP	n,adj	34	(*sublimação*, *sublimado*)
	PROP_DE_ALGO_PRODUTOR_DE	adj,n	359	(*fotógeno*, *luz*)
Purpose	FINALIDADE_DE	n,n	2,878	(*defesa*, *armadura*)
	FINALIDADE_DE_ALGO_COM_PROP	n,adj	38	(*reprodução*, *reprodutor*)
	ACCAO_FINALIDADE_DE	v,n	5,185	(*fazer_rir*, *comédia*)
	ACC_FINALIDADE_DE_ALGO_COM_PROP	v,adj	284	(*corrigir*, *correccional*)
Place	LOCAL_ORIGEM_DE	n,n	816	(*Japão*, *japonês*)
Manner	MANEIRA_POR_MEIO_DE	adv,n	1,113	(*timidamente*, *timidez*)
	MANEIRA_SEM	adv,n	121	(*devagar*, *pressa*)
	MANEIRA_SEM_ACCAO	adv,v	11	(*assiduamente*, *faltar*)
Property	PROP_DE_ALGO_REFERENTE_A	adj,n	3,520	(*dinâmico*, *movimento*)
	PROP_DO_QUE	adj,v	17,246	(*familiar*, *ser_conhecido*)

Construction of an ontology

- Handcrafting by specialists
  - Reliable, but…
  - Impracticable
  - Undesireable
  - Time-consuming
- Created (semi-) automatically
  - with the help of learning computational tools

Research Goals

- Development of computational tools capable of **learning lexico-semantic knowledge** from Portuguese text.

*A fábula é **um tipo de** narrativa que **tem o objectivo de** entreter e aconselhar.*

fábula hyponym-of *narrativa*

fábula has-pupose *entreter*

fábula has-pupose *aconselhar*

Research Goals

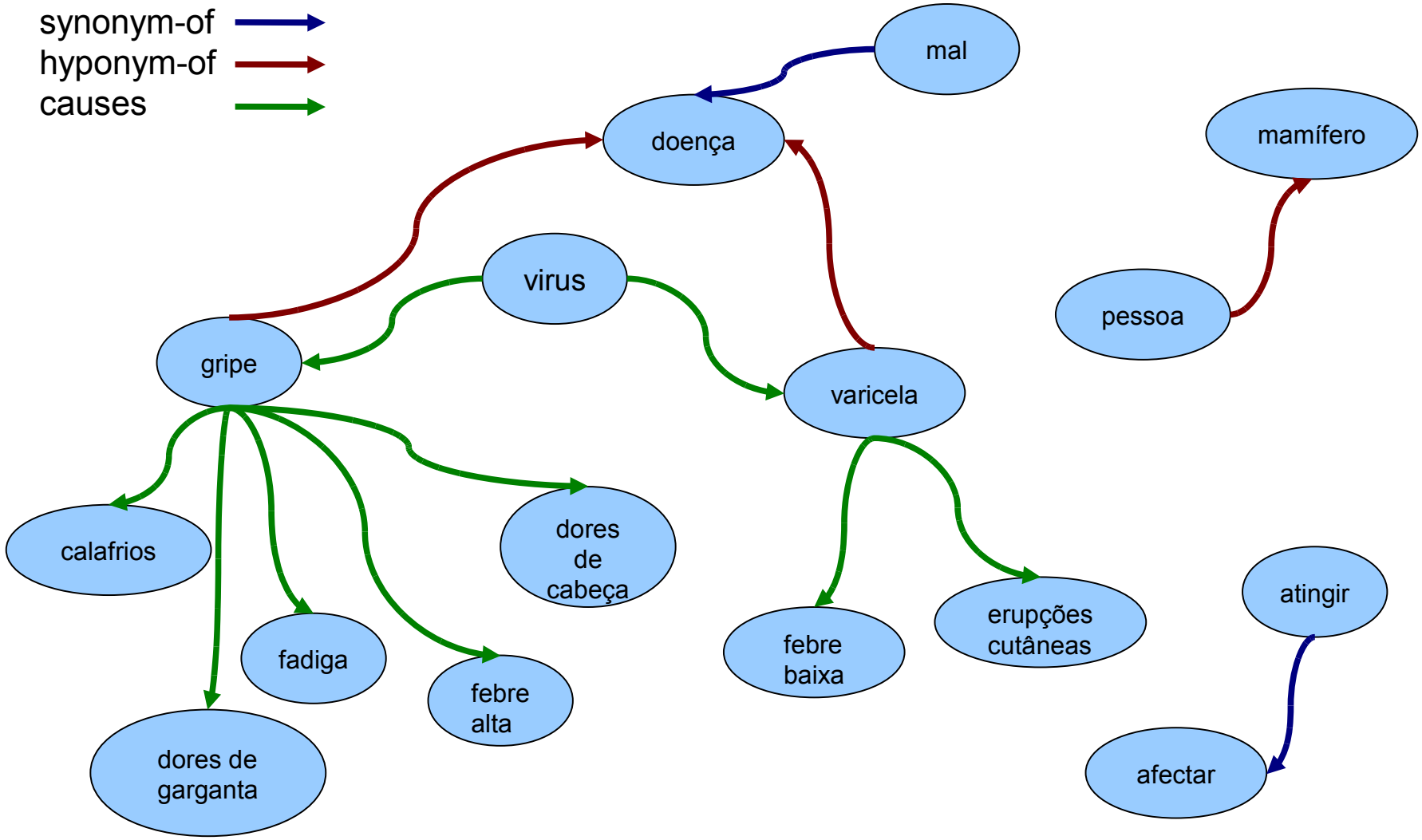
- Construction of **Onto.PT**, a freely available lexical ontology for Portuguese, created semi-automatically.
- Several evaluation methodologies, with special focus to **semi-automatic** evaluation.

Motivation

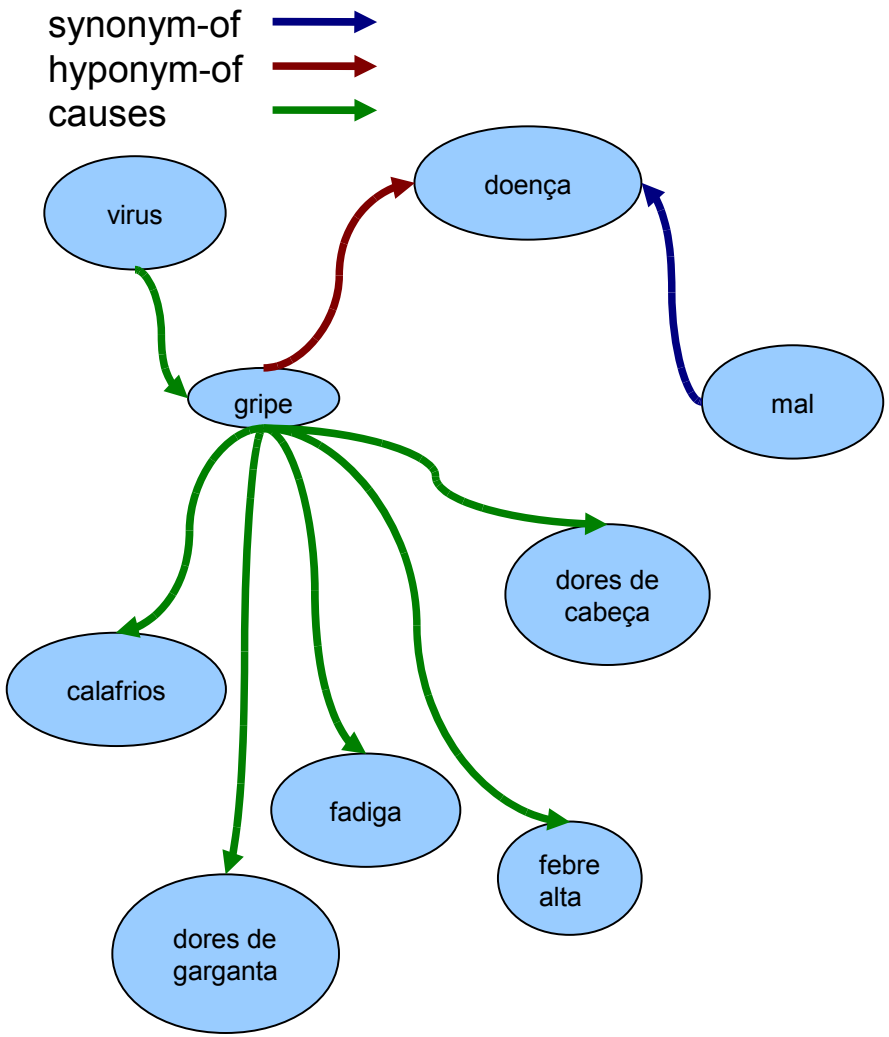
- Lexical ontologies are useful for many NLP tasks, such as Information Retrieval
  - Which of these snippets are related?

Snippet A	A gripe é causada por um vírus altamente contagioso que afecta aves e mamíferos. Tipicamente, a gripe é transmitida por mamíferos infectados por meio do ar e por aves infectadas por meio de suas secreções.
Snippet B	A varicela é causada por um vírus altamente contagioso que afecta essencialmente crianças. Tipicamente, a varicela é transmitida através da inalação de gotículas presentes no ar, que contêm o vírus.
Snippet C	O mal pode facilmente atingir várias pessoas e os seus principais sintomas são calafrios, febre alta, dores de garganta, dores de cabeça e fadiga.

Motivation


synonym-of
hyponym-of
causes
mal
doença
mamífero
virus
pessoa
gripe
varicela
calafrios
dores de cabeça
fadiga
febre alta
dores de garganta
febre baixa
erupções cutâneas
atingir
afectar

Motivation



- Some conclusions:
  - All snippets are about diseases (*doenças*)
  - Snippets A and C are related to flu (*gripe*)
- Questions can be answered:

Q: O que é a gripe?
*A: Uma **doença** ou **mal**, transmitida por um **vírus**, que provoca **calafrios**, **febre alta**, **dores de cabeca**, **dores de garganta** e **fadiga**.*

Approach

- Learning from MRDs (machine readable dictionaries)
  - Semantic authorities
  - Restricted vocabulary
  - Already structured on words and meanings
  - General knowledge
- Learning from unrestricted text
  - Much available
  - Rich on specific domains
  - Unrestricted text

Extraction of Lexical Knowledge from MRDs

- Some historical remarks:
  - Calzolari (1980, 1982, 1984) and Amsler (1981) explored the structure of MRDs in order to extract lexical information from them.
  - Chodorow et al. (1985) developed procedures capable of extracting *tangled hierarchies*.
  - Alshawi (1987, 1989) developed semantic grammars for one dictionary.
  - MindNet (Richardson et al. (1998)) is a lexical knowledge base extracted semi-automatically from MRDs.

Extraction of Lexical Knowledge from MRDs

- Use PAPEL as a starting point
- Adapt the methodology to other MRDs
- Merge the results adequately

PAPEL
Extraction

- Extraction procedure
  - PEN* parser + semantic grammars

1

cometa, s. m.
astro geralmente constituído
por núcleo, cabeleira e cauda

2

```
[RAIZ]
   [QUALQUERCOISA]
     > [astro]
     [QUALQUERCOISA]
       > [geralmente]
   [PADRAO_CONSTITUIDO]
     [VERBO_PARTE_PP]
       > [constituído]
     [PREP]
       > [por]
   [ENUM_PARTE]
     [PARTE_DE]
       > [núcleo]
     [VIRG]
       > [,]
     [ENUM_PARTE]
       [PARTE_DE]
         > [cabeleira]
       [CONJ]
         > [e]
       [PARTE_DE]
         > [cauda]
```

3

núcleo PARTE_DE cometa
cabeleira PARTE_DE cometa
cauda PARTE_DE cometa

*available through http://code.google.com/p/pen/

PAPEL

Post-processing

- All relations converted to their direct type
  - *manga* INCLUI *punho* >> *punho* PARTE_DE *manga*
  - *dor* RESULTADO_DE *distensão* >> *distensão* CAUSADOR_DE *dor*
- Lemmatization of arguments
- Correction of the relation name
  - *loucura* ACCAO_QUE_CAUSA *desvario* >> *loucura* CAUSADOR_DE *desvario*

Ontology Learning from Textual Corpora

- Enrich knowledge extracted from MRDs.
- Discovery of patterns (Hearst 1992)
  - Choose 2 related terms (e.g. *dog* and *animal*)
  - Look for patterns occurring between them
- Application of the patterns to corpora
  - Development of semantic grammars
  - Extraction of relations between terms
- Using this procedure to extract several relations.

Ontology Learning from Textual Corpora

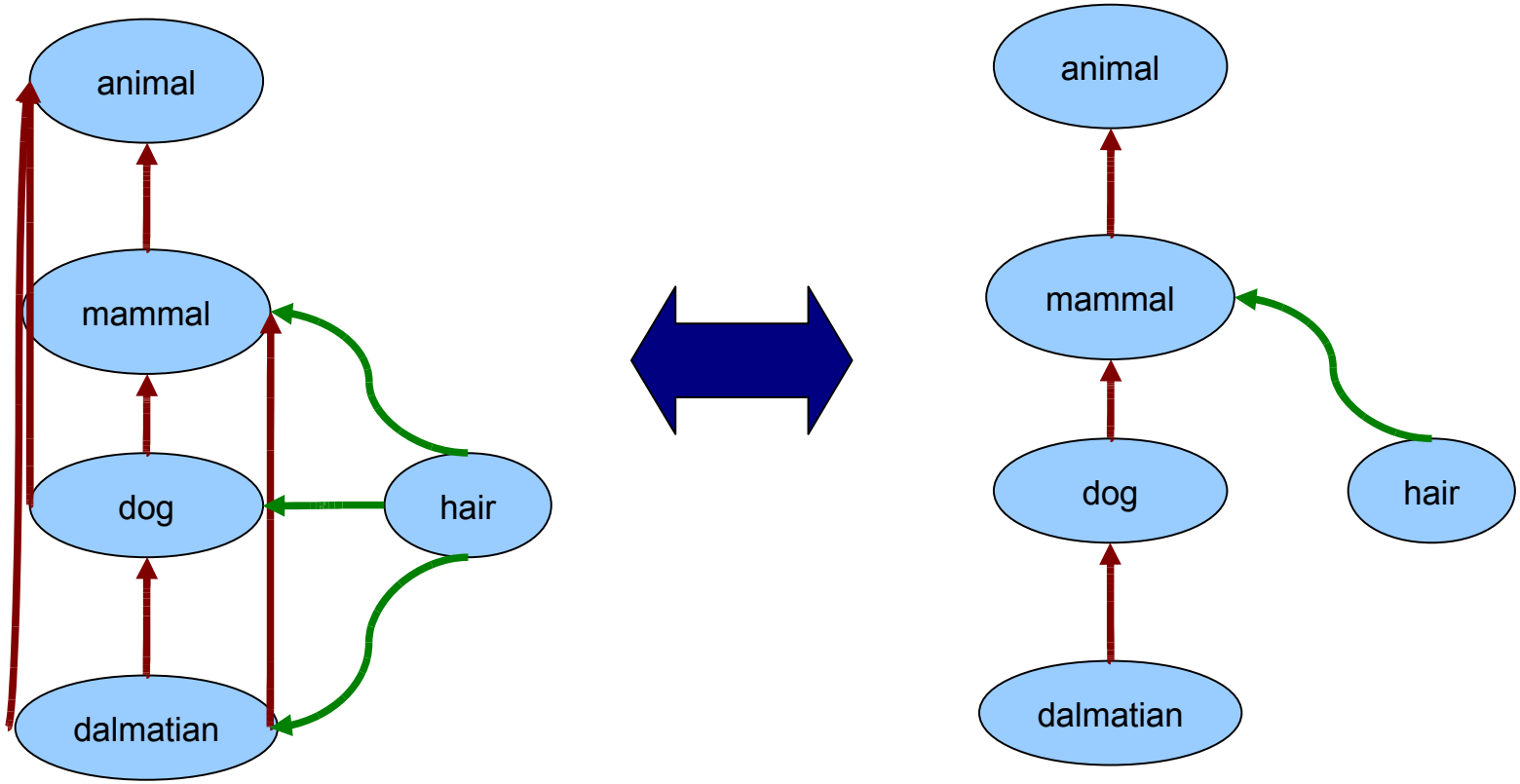
- Improve precision
  - Similarity metrics, based on the co-occurrence of terms (Caraballo 1999; Cederberg & Widdows 2003)
    - e.g. discard hypernym relations between non-similar terms
  - Take advantage of syntactical annotation.
- Improve recall
  - Take advantage of specific constructions (Cederberg & Widdows 2003)
    - "*..., dogs, cats, parrots and rabbits...*" may suggest that all of these have a common hypernym.

Resource construction

- Organisation of words and their senses
  - Association of terms (Roark and Charniak (1998); Pantel and Lin (2002))
  - Exploitation of the (ambiguous) network structure.
  - Co-occurrence in example sentences
- Integration of new knowledge

Resource construction

- Knowledge “compression”
- Knowledge discovery



Evaluation

- Inspired by existing methods to evaluate domain ontologies (Brank et al. 2005)
  - Manual
  - Comparison with a *golden standard*
  - Comparison with a collection of documents about a domain covered by the ontology
  - Task-based

Evaluation

- Adapted to lexical ontologies
  - Manual will be similar
  - Comparison with a *golden standard*, requires a semantic resource for Portuguese
    - Tep can be used to evaluate synonymy in Onto.PT
    - Corpora with semantic annotation to calculate the precision and recall of the tools.

$$Precision = \frac{Correct\ answers}{Given\ answers}$$

$$Recall = \frac{Correct\ answers}{Possible\ answers}$$

Evaluation

- Adapted to lexical ontologies
  - Lexical coverage (Demetriou and Atwell 2001)
    - *How many words of a corpus are in the ontology?*
  - Task-based
    - Using the ontology with other application (e.g. Q&A system, text generator ...)
  - Alternatives
    - Search in corpora for patterns than indicate each relation (Etzioni et al. 2005, Gonçalo Oliveira et al. 2009)

PAPEL
Evaluation of synonymy

- TeP 2.0 as a golden resource
- Relations with terms that are not both in PAPEL 1.0 and TeP 2.0 are removed
  - 50% of PAPEL in TeP, 39% of TeP in PAPEL
- Expansion: (A SINONIMO_DE B) e (B SINONIMO_DE C) >> (A SINONIMO_DE C)
  - 19% of PAPEL in TeP, 90% of TeP in PAPEL
  - Incorrections
    - A=*ruína*, B=*queda*, C=*habilidade*
    - >> *ruína* SINONIMO_DE *habilidade*

PAPEL

Evaluation of other (noun to noun) relations

- Relations rendered as natural language patterns
- Searched in CETEMPúblico corpus

Relation	Correct	Support
língua HIPERONIMO_DE *italiano*	Yes	*As **línguas latinas**, **como o italiano** ou o português, tornam-se mais fáceis por causa das vogais.*
arbusto PARTE_DE *floresta*	Yes	*A **floresta** é um conjunto de árvores, **arbustos** e ervas de várias qualidades e tamanhos.*
cólera CAUSADOR_DE *diarreia*	Yes	*A **cólera provoca fortes diarreias** e vómitos e pode levar à desidratação e, consequentemente, à morte em poucas horas.*
oliveira PRODUTOR_DE *azeitona*	Yes	*Também a quantidade e tamanho das **azeitonas produzidas por uma oliveira** biológica é inferior, já que não são utilizados compostos de azoto que ajudam a planta a crescer.*
recrutamento FINALIDADE_DE *inspecção*	Yes	*Menos de metade dos jovens entre os 20 e os 22 anos apresentaram-se às **inspecções para recrutamento**, revelou o ministro da Defesa.*
músico PARTE_DE *música*	No	*... um espectáculo baseado na obra "Cantos de Maldoror", de Lautréamont, com **música composta pelo músico** inglês Steven Severin...*
fim FINALIDADE_DE *sempre*	No	*Sicília aponta **sempre para o fim** do dia, para o fim da luz.*

PAPEL

Evaluation of other (noun to noun) relations

- Results for PAPEL 1.0 + CETEMPúblico 1.7

Relation	Relations w/ args in CETEMPúblico	%	Sample	%	Hits	%
Hypernymy	40,079	63%	3,145	8%	560	18%
Meronymy	3,746	35%	2,343	63%	521	22%
Causation	557	50%	557	100%	20	4%
Producer	414	44%	414	100%	12	3%
Purpose	1,718	59%	1,718	100%	173	10%

- Some correct relations are not found:
  - fruto HIPERONIMO_DE alperce
  - algoritmia PARTE_DE matemática
  - ausência CAUSADOR_DE saudade
  - aquecimento FINALIDADE_DE salamandra
  - ...

Concluding remarks

- Important contributions for Portuguese NLP are expected:
  - **Onto.PT**, a free lexical ontology for Portuguese
  - Computional tools for semi-automatic extraction of semantic knowledge from Portuguese text
  - Methodologies to evaluate lexical ontologies
  - Scientific papers
  - PhD thesis
- Current work available through:
  - http://eden.dei.uc.pt/~hroliv/phd.html

Ontology Learning for Portuguese

Thank you!

Hugo Gonçalo Oliveira
hroliv@dei.uc.pt

University of Coimbra
Faculty of Sciences and Technology
Department of Informatics Engineering

Knowledge and Intelligent Systems Laboratory
Cognitive and Media Systems Group
Centre of Informatics and Systems of the University of Coimbra