Automatic Extraction and Validation of Lexical Ontologies from text

Hernani Costa

hpcosta@student.dei.uc.pt

Cognitive & Media Systems Group
CISUC, University of Coimbra

Coimbra, September, 2010

Understanding the meaning of natural language

- For making people and machines communicate
  - tools capable of exchanging well-defined and unambiguous information
  - manipulation of natural language
  - encoding it into a formal language

- Attempts to formalise semantic knowledge in a kind of lexical ontology (Princeton WordNet (Fellbaum (1998)))
- Similar resources for Portuguese (WordNet.BR (Dias-da-Silva (2006))), WordNet.PT (Marrafa et al. (2006)))

Introduction

- Knowledge bases are useful resources for NLP, however...
- Their creation and maintenance involves intensive human effort
- Automatic creation/enrichment from textual resources is an alternative
  - Higher coverage, easier update, but...
  - Precision is lower
  - **Evaluation requires once again intensive human labour!**

Information extraction (IE)

Automatic extraction of structured information from natural language inputs.

- "A car **is a** vehicle that **has an** engine and **aims to** move planets."
  - *vehicle* HYPERNYM_OF *car*
  - *engine* PART_OF *car*
  - *car* PURPOSE_OF *move planets*

How to automatically validate semantic knowledge?

- "A car **is a** vehicle that **has an** engine and **aims to** move planets."
  - √ *vehicle* HYPERNYM_OF *car*
  - √ *engine* PART_OF *car*
  - X ~~*car* PURPOSE_OF *move planets*~~

Information retrieval (IR)

Locating specific information in natural language resouces.

- Approaches based on the occurrence of words in documents
- Distributional similarity metrics
  - Corpus Distributional Metrics
    - ★ Cocitation (Small (1973))
    - ★ LSA (Deerwester et al. (1990))
    - ★ PMI-IR (Turney (2001))
    - ★ ...
  - Web Distributional Metrics (Bollegala et al. (2007))
    - ★ WebJaccard
    - ★ WebOverlap
    - ★ WebDice
    - ★ ...

Goals

1. Discovery of new lexico-syntactic patterns (automatically and by observation)

2. System capable of:
   - extract written data from textual resources
   - extract semantic information from unstructured text
   - infer new knowledge based on compound nouns
   - validate and evaluate semantic knowledge

   $\Downarrow$

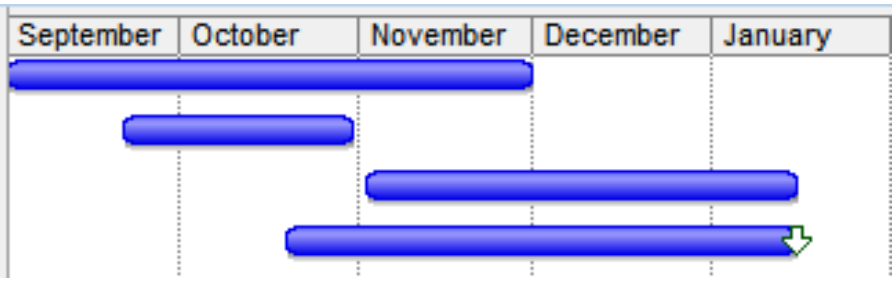
   Hybrid system (linguistic + statistic)

3. Compare knowledge-bases

Research planning

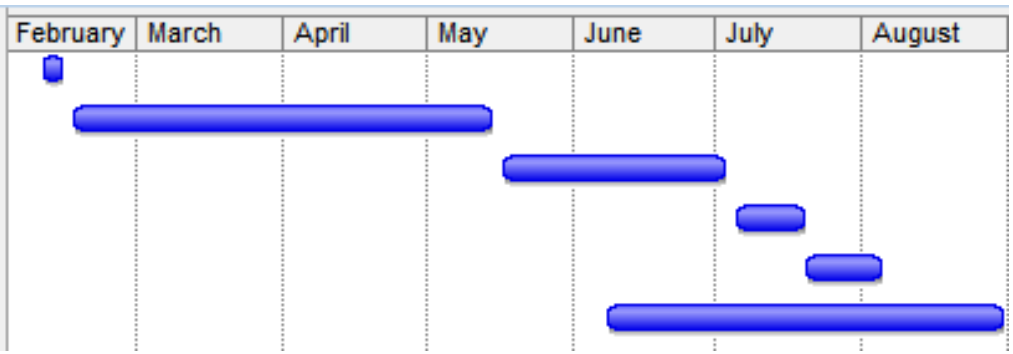
1^{st} semester

	Task Name	Duration
1	Bibliography Revision	65 days
2	Discovery of patterns	30 days
3	First System Prototype	55 days
4	Thesis Proposal	65 days

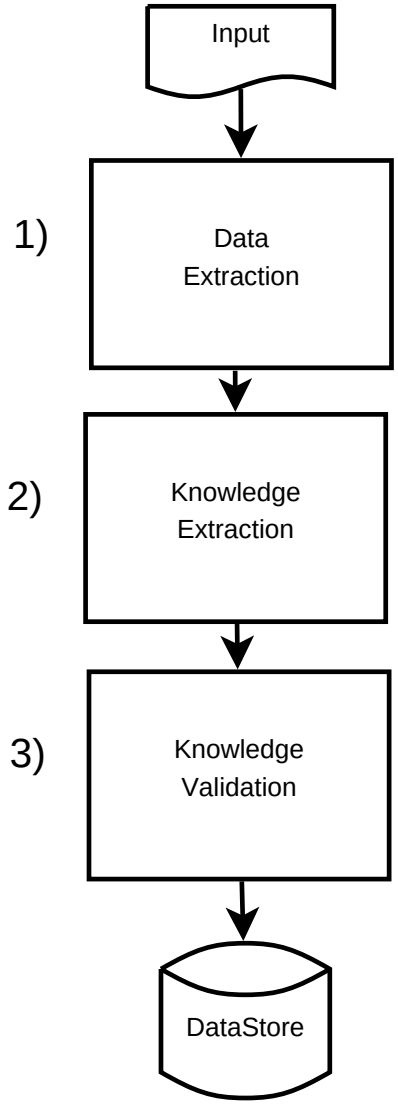


2^{nd} semester

	Task Name	Duration
1	Process Diagram Elaboration	4 days
2	First System Prototype	65 days
3	Second System Prototype	35 days
4	Studying System Improvements	11 days
5	Comparing Knowledge-bases	12 days
6	Final Thesis Elaboration	60 days

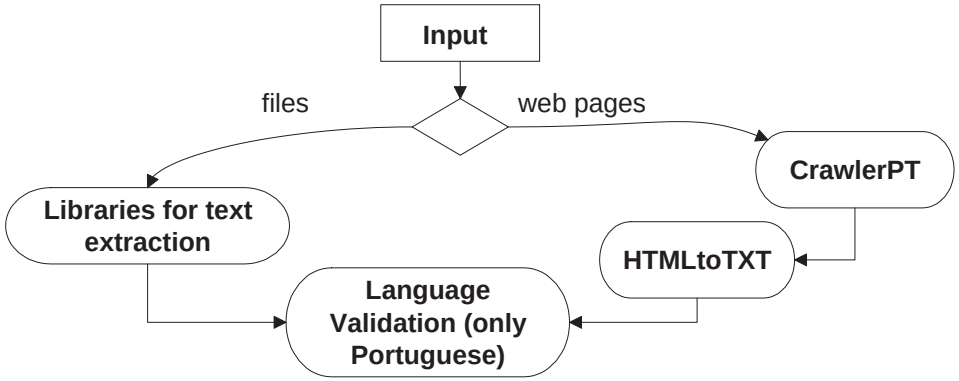


System Architecture



System Architecture

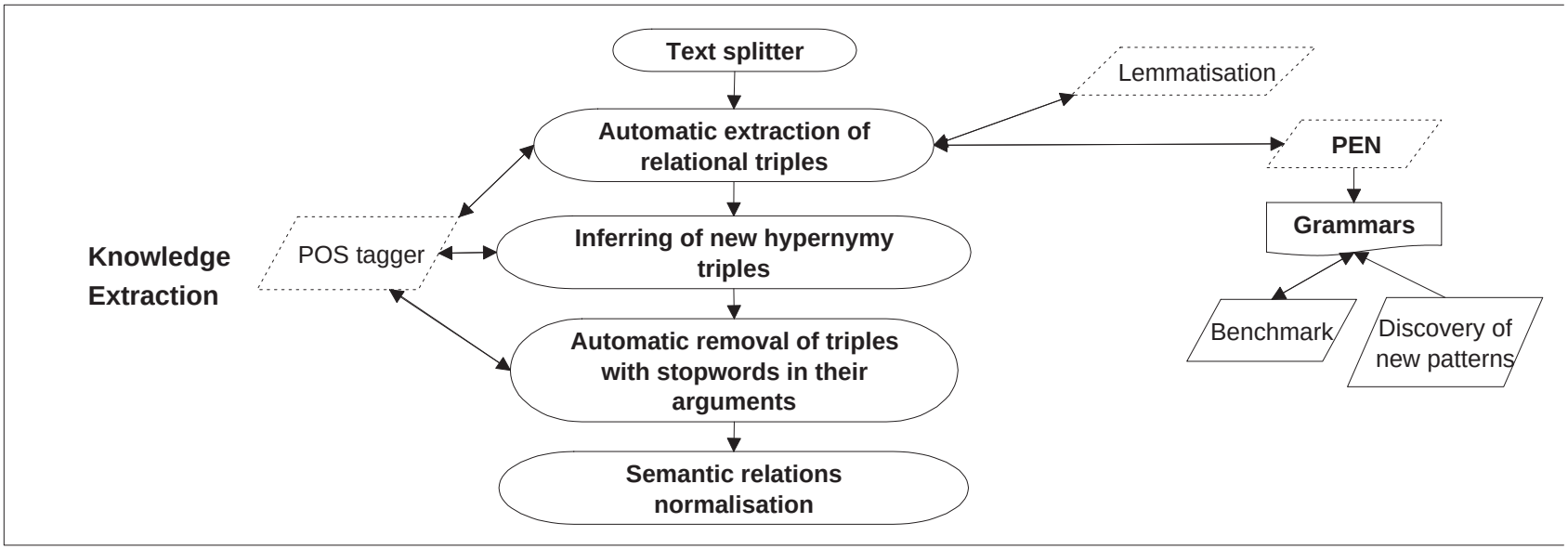
1) Data Extraction modules



- Extract written data from different textual resources
  - Docs, pdfs, rdf, txts, ...
  - Crawl data from the Web
- Only Portuguese data is considered

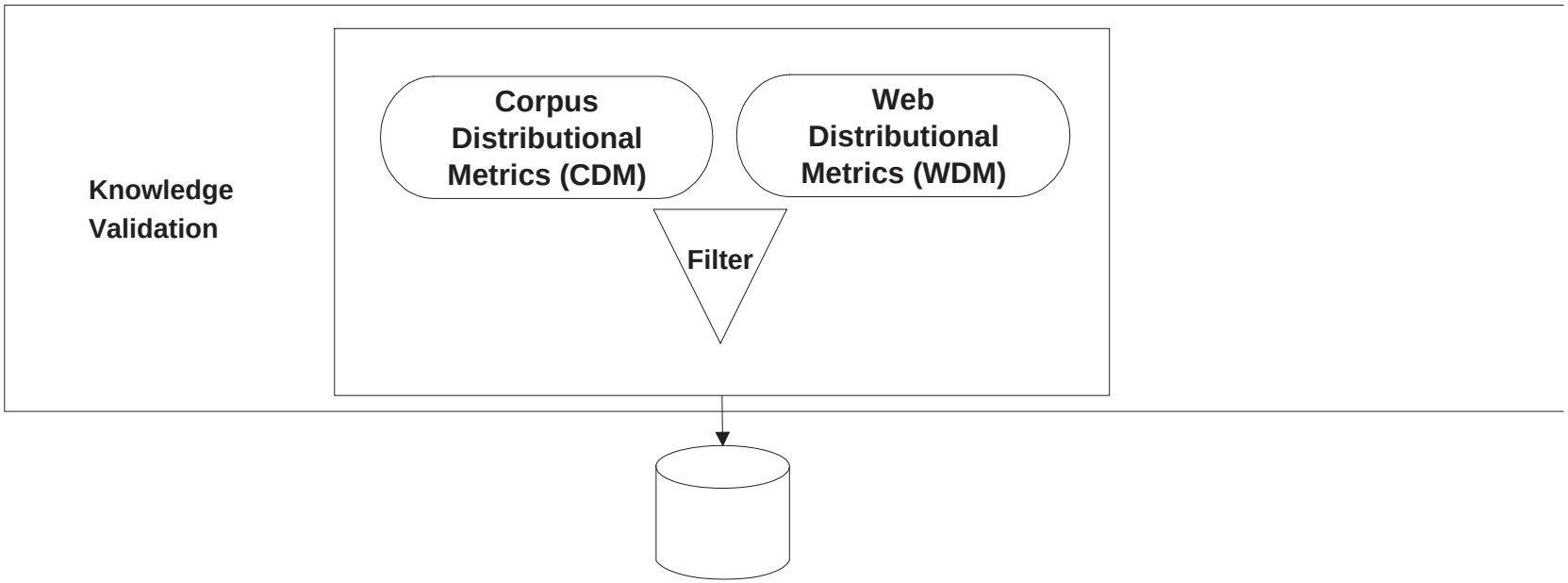
System Architecture

2) Knowledge Extraction



System Architecture

3) Knowledge Validation

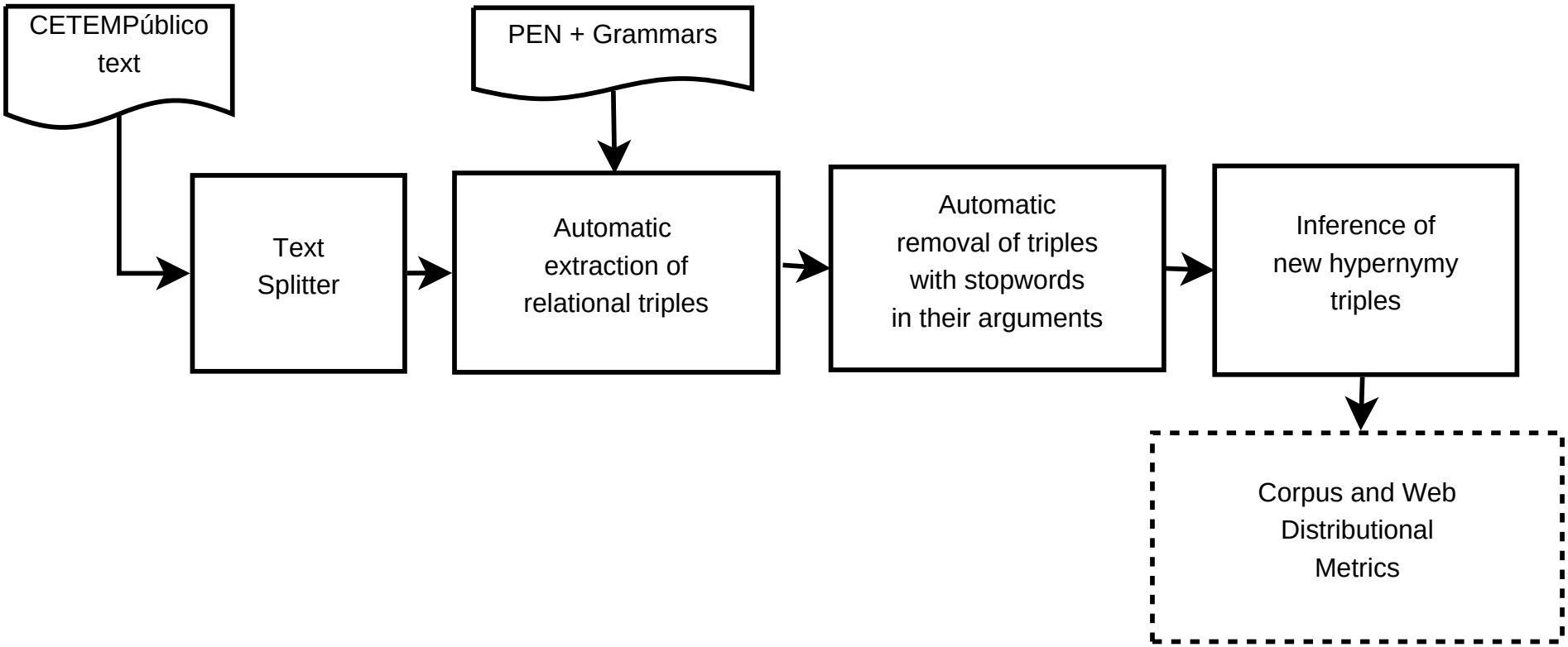


Experimental Work

1. Knowledge extraction from CETEMPúblico
2. Knowledge extraction from Wikipedia abstracts
3. Comparing prototype 1 to prototype 2
4. Knowledge-bases comparison

Experiment 1: knowledge extraction from CETEMPúblico

System modules



Experiment 1: knowledge extraction from CETEMPúblico

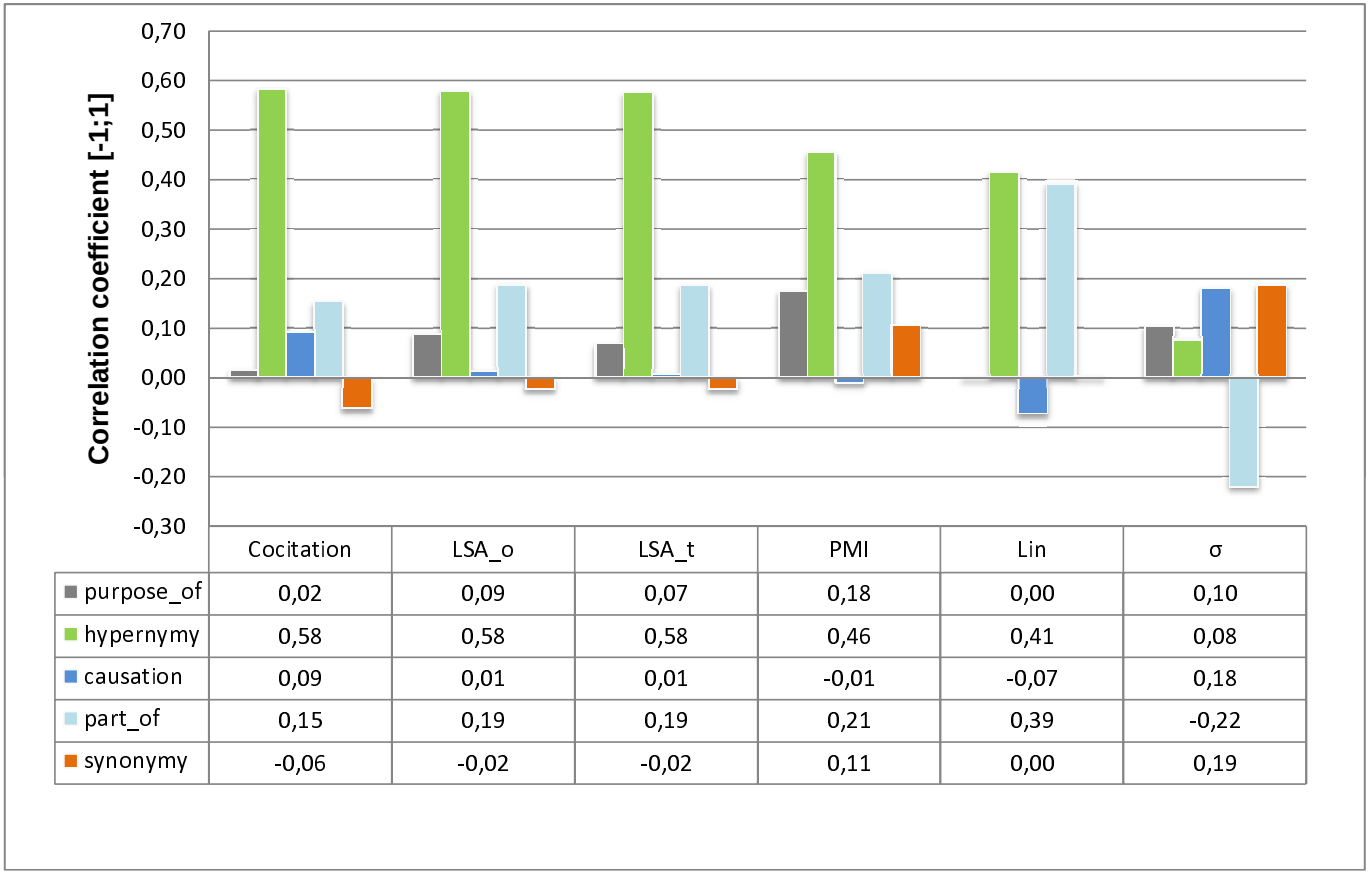
Set-up and Results

- CETEMPúblico[1] (Santos and Rocha (2001)) corpus, the annotated version
  - 28,000 documents
  - 30,100 unique content words (nouns, verbs and adjectives)
  - *term-document* matrix
  - *term-term* matrix

- Triples obtained
  - extracted: 20,308
  - discarded: 5,844
  - inferred: 2,492
  - final triple set: **16,956**

[1] http://www.linguateca.pt/cetempublico

Manual Evaluation vs. Corpus Distributional Metrics



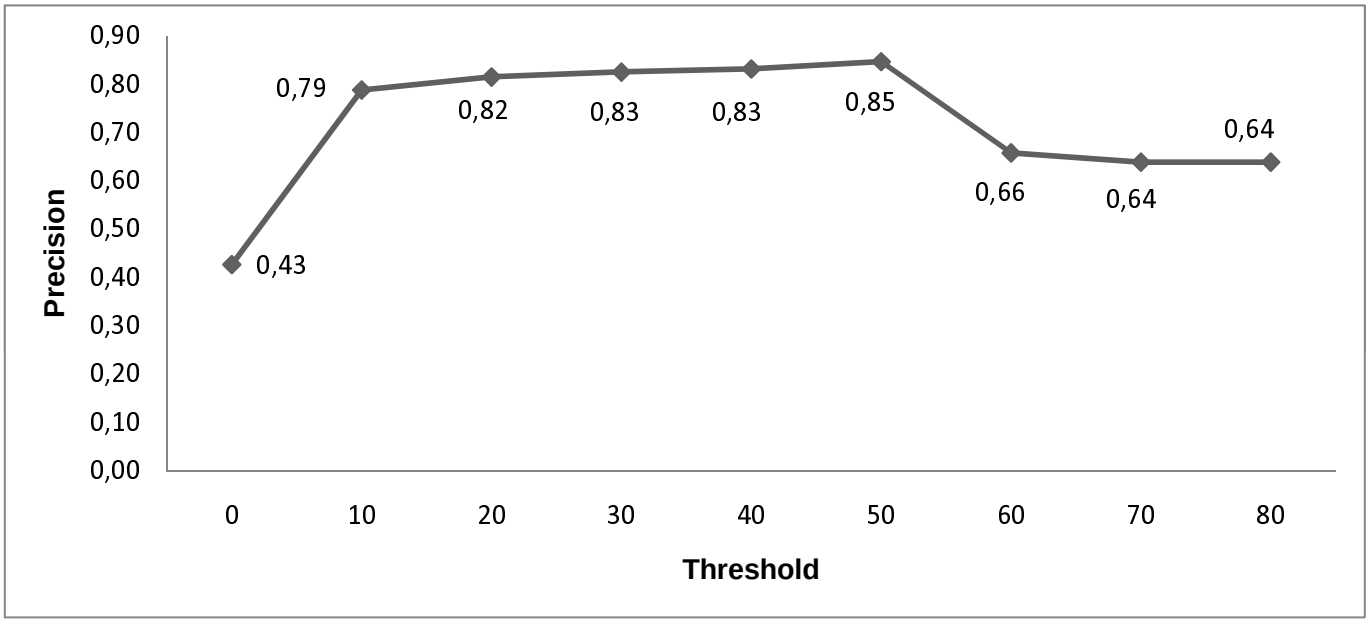
	Cocitation	LSA_o	LSA_t	PMI	Lin	σ
purpose_of	0,02	0,09	0,07	0,18	0,00	0,10
hypernymy	0,58	0,58	0,58	0,46	0,41	0,08
causation	0,09	0,01	0,01	-0,01	-0,07	0,18
part_of	0,15	0,19	0,19	0,21	0,39	-0,22
synonymy	-0,06	-0,02	-0,02	0,11	0,00	0,19

- *term-document* matrix statistically dominates *term-term* matrix on 89%
- *term-term* matrix statistically dominates *term-document* matrix on 72%

Metrics-based threshold

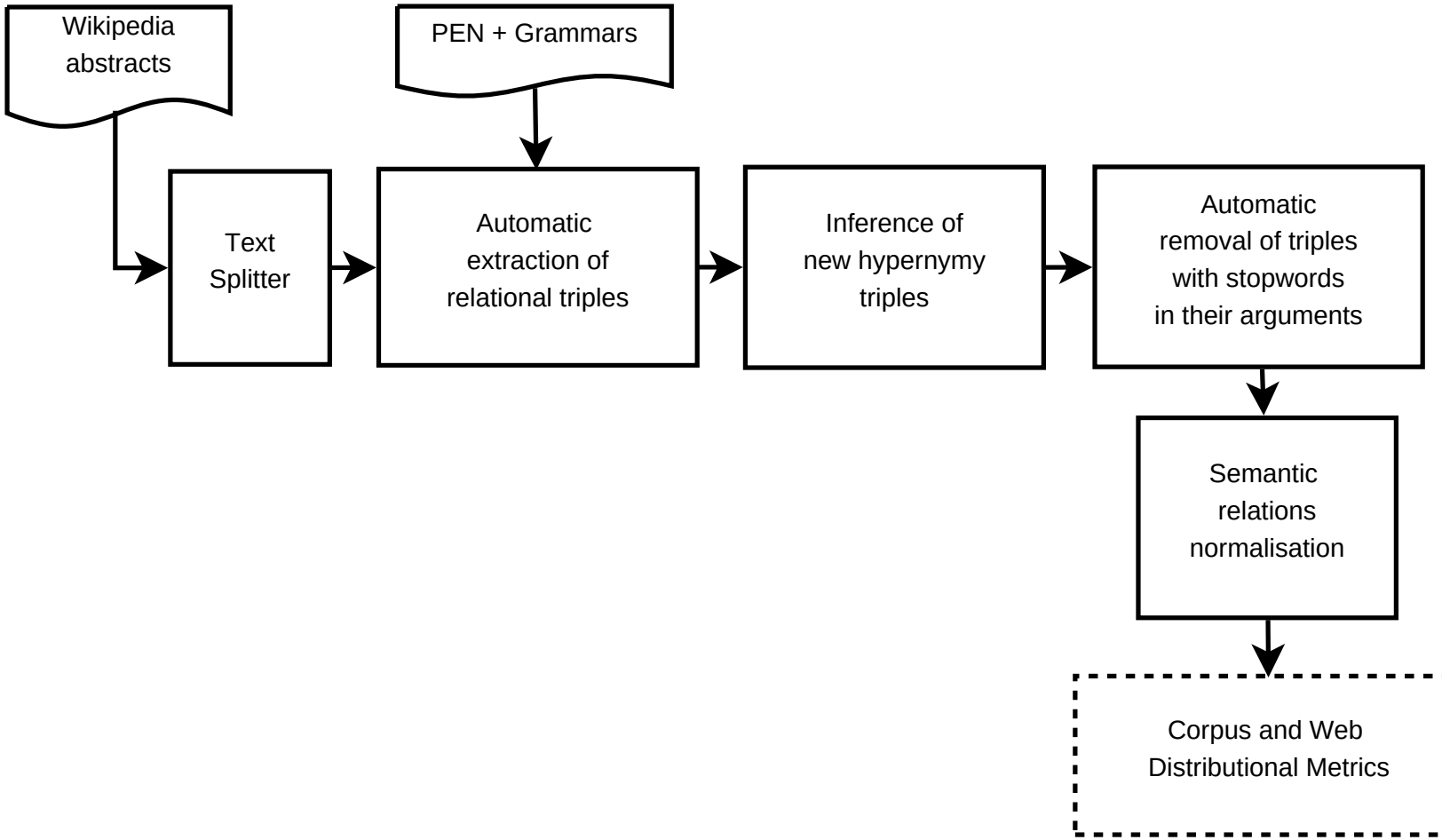
Increasing the threshold for hypernymy relation

- Threshold based on the Cocitation value
- Increased gradually for **hypernymy** triples
- 50 seems to be a good cut-point



Experiment 2: knowledge extraction - Wikipedia abstracts

System modules



Experiment 2: knowledge extraction - Wikipedia abstracts

Set-up and Results

- Wikipedia abstracts
  - 37,898 sentences
  - without named entities
- Triples obtained
  - extracted + inferred: 70,150
  - discarded: 9,947
  - final triple set: **60,203**

Studing Patterns Efficiency

Table: Quantity of triples extracted based on their indicative patterns.

Relation	Pattern	Evaluated			
		3	**2**	**1**	**0**
Hypernymy	*multi-word term*	72	7	75	32
	é uma espécie de	54	96	0	0
	é um|uma	87	11	0	15
	é um género de	24	0	0	0
Synonymy	ou	154	2	0	2
	também conhecido|a|os|as por|como	60	4	0	4
Part_of	inclui|incluem	34	0	2	15
	grupo de	17	3	1	0
Purpose	utilizado|a|os|as para|como|em|no|na	71	16	1	20
	usado|a|os|as para|como|em|no|na	41	3	1	4
Causation	causado|a|os|as	27	11	1	10

- Caption:

  3 $\Rightarrow$ correct **2** $\Rightarrow$ contains strange entities

  1 $\Rightarrow$ too general or specific **0** $\Rightarrow$ incorrect

Experiment 3: comparing prototype 1 to prototype 2

Set-up and Results

- System prototype 2 on CETEMPúblico
  - studying the system improvements

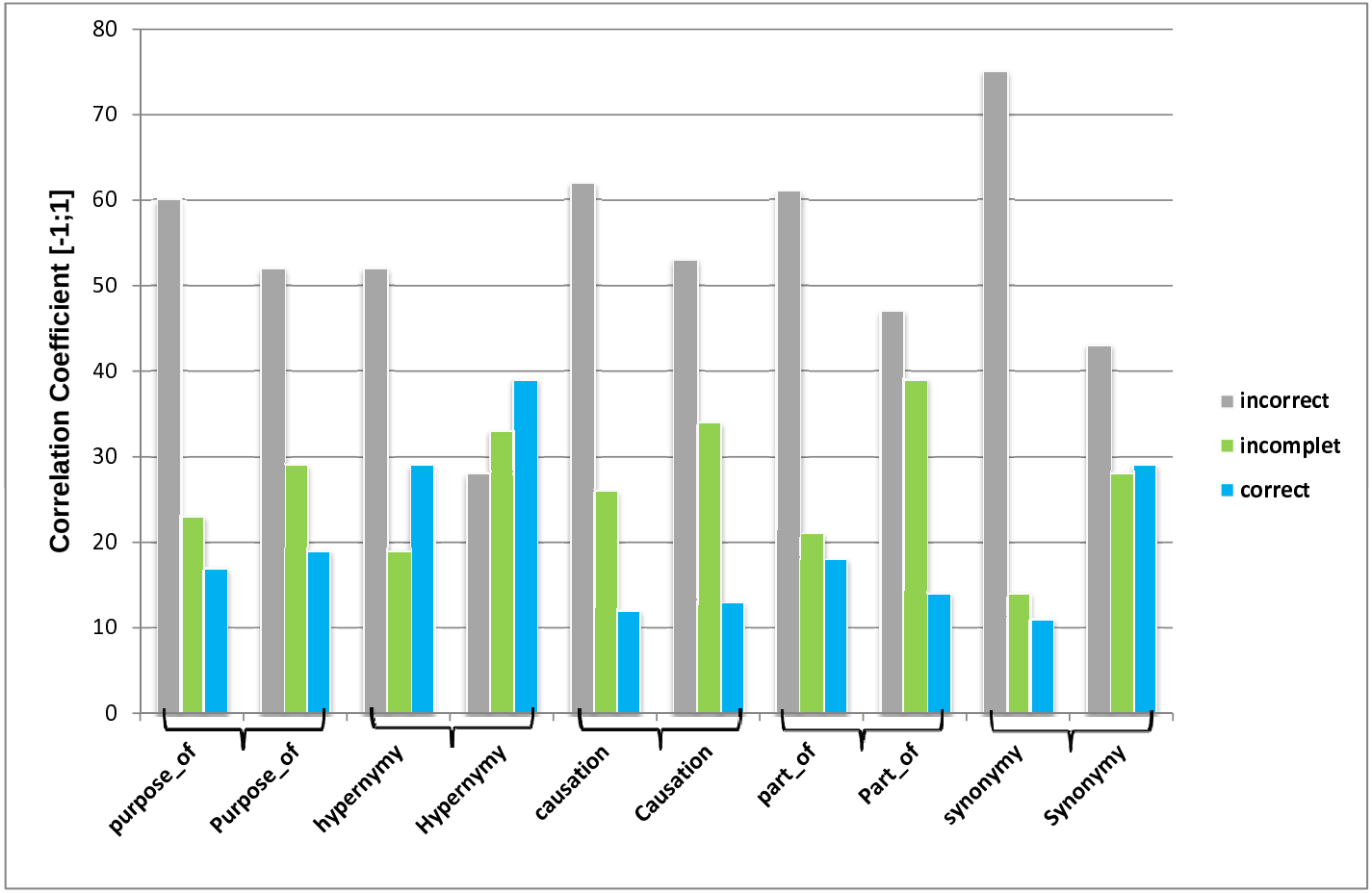
- Number of triples extracted from the CETEMPúblico corpus:

Table: *Experiment 3* vs. *Experiment 1*.

Relation	Experiment 3	Experiment 1
Hypernymy	286,960	9,365
Causation	3,037	2,660
Purpose	3,779	3,288
Part_of	1,759	1,373
Synonymy	254	270
TOTAL	*295,789*	*16,956*

Experiment 3: comparing prototype 1 to prototype 2

Manual evaluation (first vs. second approach) percentages



Caption:

- *Experiment 1* values → relation name starts with lowercase letter
- *Experiment 3* values → relation name starts with Uppercase letter

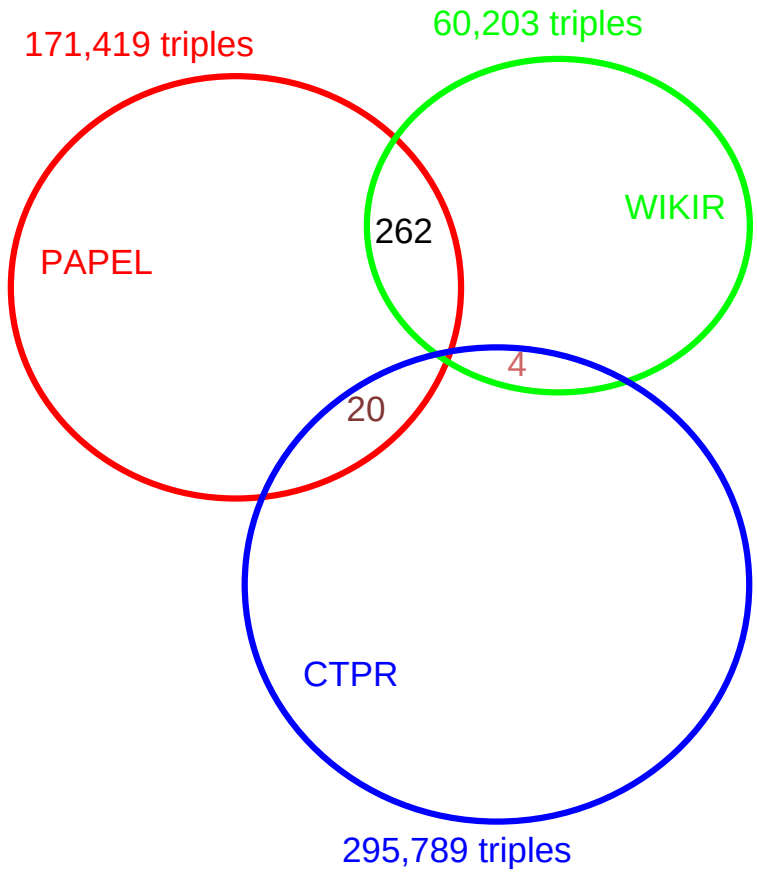
Experiment 4: knowledge-bases comparison

Set-up

- CTPR → knowledge extracted from *Experiment 3*
- WIKIR → knowledge extracted from *Experiment 2*
- PAPEL → knowledge extracted from a dictionary (Gonçalo Oliveira et al. (2009))

Experiment 4: knowledge-bases comparison

- WIKIR: associated to the world and human knowledge
- CTPR: specific knowledge
- PAPEL: knowledge about the words and their meanings



- Common knowledge = C1 + C2
  - C1 → common triples
  - C2 → common triples - but with different relation

Contributions

1. Modules capable of interpreting text contained in different documents
2. New indicative patterns to the semantic relations covered by our system (hypernymy, synonymy, part_of, purpose_of and causation)
3. Method to infer hypernymy relations from compound nouns
4. IR metrics applied to IE
5. Automatic evaluation proposal (Web + lexico-syntactic patterns)
6. Method to compare knowledge-bases

Publications

- ECAI[2] 2010, workshop LaTeCH[3] 2010
  - Costa et al. (2010) (available through `http://student.dei.uc.pt/~hpcosta/papers/ecai2010.pdf`)
- INForum[4] 2010
  - Gonçalo Oliveira et al. (2010) (available through `http://student.dei.uc.pt/~hpcosta/papers/inforum2010.pdf`)

[2] http://ecai2010.appia.pt
[3] http://ilk.uvt.nl/LaTeCH2010
[4] http://inforum.org.pt/INForum2010

Future Work

Besides more experimentations, also more ideas can be explored:

- **Discovery on new semantic patterns**
  - using a bigger corpus, such as the Web
- **Extract semantic knowledge using machine learning techniques**
  - more versatile as regards the variations in lexico-syntactic patterns
- **Studying the better windows size**
  - to understand how it influence the corpus distributional metrics results
- **Weighting triples**
  - using external resources to assign weights to the triples, or
  - weight the entities based on their occurrence in some textual resource
- **Evaluation module**
  - it would be interesting their deeper study

References I

Bollegala, D., Matsuo, Y., and Ishizuka, M. (2007). *Measuring semantic similarity between words using web search engines*, pages 757–766. ACM Press, Proc. 16th International Conference on World Wide Web (WWW'07) edition.

Costa, H., Gonçalo Oliveira, H., and Gomes, P. (2010). The Impact of Distributional Metrics in the Quality of Relational Triples. In *Proc. ECAI 2010, Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH'10)*.

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.

Dias-da-Silva, B. (2006). Wordnet.Br: an exercise of human language technology research. In *Proc. 3rd International WordNet Conference (GWC'06)*, pages 22–26, Jeju Island, Korea.

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

Gonçalo Oliveira, H., Costa, H., and Gomes, P. (2010). Extracção de conhecimento léxico-semântico a partir de resumos da Wikipédia. In *Proc. 2nd INFORUM 2010 Workshop on Gestão e Tratamento de Informação (INFORUM'10)*.

Gonçalo Oliveira, H., Santos, D., and Gomes, P. (2009). Relations extracted from a Portuguese dictionary: results and first evaluation. In *Local Proceedings of the 14th Portuguese Conference on Artificial Intelligence (EPIA'09)*.

References II

Marrafa, P., Amaro, R., Chaves, R. P., Lourosa, S., Martins, C., and Mendes, S. (2006). WordNet.PT new directions. In Sojka, P., Choi, K., Fellbaum, C., and Vossen, P., editors, *Proc. 3rd International WordNet Conference (GWC'06)*, pages 319–320.

Santos, D. and Rocha, P. (2001). Evaluating CETEMPúblico, a free resource for portuguese. In *Proc. 39th Annual Meeting on Association for Computational Linguistics (ACL'01)*, pages 450–457, Morristown, NJ, USA. ACL.

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 24(4):265–269.

Turney, P. D. (2001). Mining the Web for Synonyms: PMI–IR versus LSA on TOEFL. In Raedt, L. D. and Flach, P., editors, *Proc. 12th European Conference on Machine Learning (ECML'01)*, volume 2167, pages 491–502. Springer.

Thank you!