

Building a Large Scale Lexical Ontology for Portuguese

Nuno Seco

Linguatca Node of Coimbra

<http://linguataca.dei.uc.pt>

Agenda

- Motivations
- Goals
 - Ontology Extraction
 - Ontology Evaluation
 - Study the Systematicity of Polysemy in the Lexicon using the ontology.
- What has been done so far...

Motivation

- Communication (in natural language) is a knowledge hungry task.
 - Grammatical knowledge (e.g., SVO, VSO, ...)
 - Cultural knowledge
 - Common sense knowledge
- If computers are to do NLP they need knowledge.

Motivation

- Some properties complicate the automatic processing:
 - Metaphorical nature
 - Context dependent
 - Vagueness
 - Creative
 - Diachronic
- ... but these properties are the result of human usage. and makes language use easy by humans!

Motivation

- So what we need is a resource* that can be used by a machine and makes explicit the effect of these properties.

A Lexical Ontology for Portuguese

* Be aware as this is only a snapshot of the language in a particular point in time.

Motivation

- Two strategies are usually followed:
 - Manual construction
 - WordNet
 - Cyc
 - HowNet
 - (Semi) Automatic construction ←
 - MindNet
 - KnowItAll
 - PAPEL (**P**alavras **A**ssociadas **P**orto **E**ditora **L**inguateca)

Motivation

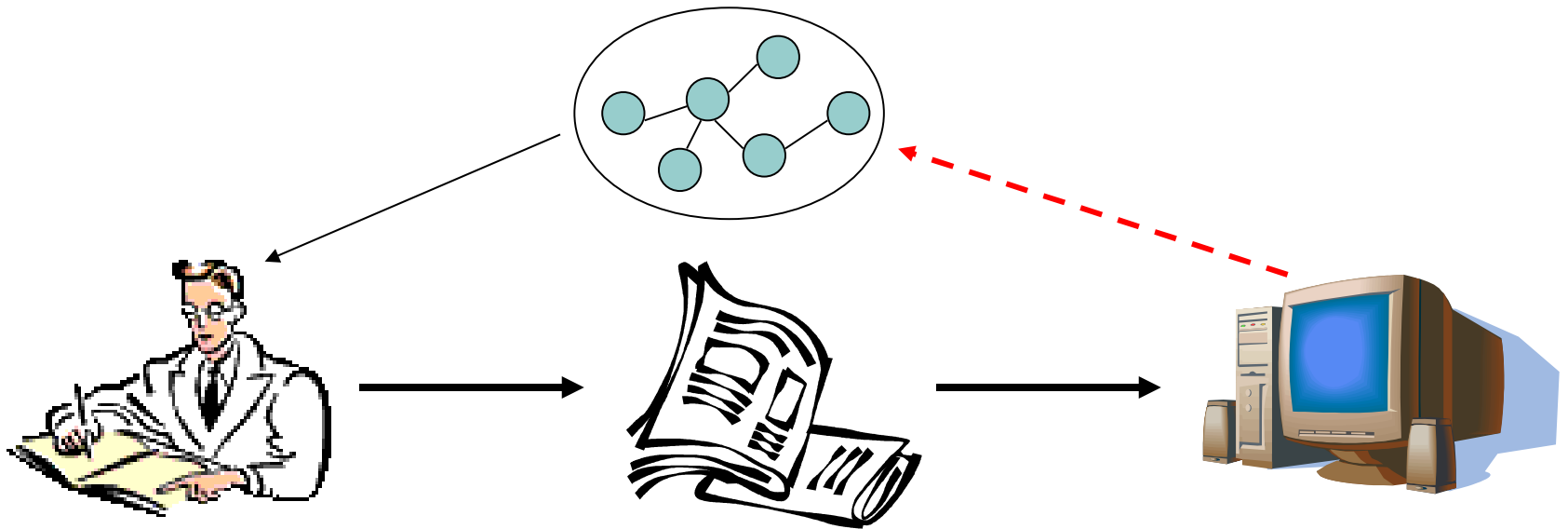
- So what can be done with a lexical ontology?
 - Information Retrieval
 - Machine Translation
 - Question Answering
 - Semantic Similarity Judgments
 - Concept Creation / Explanation

Goals

- Extract the semantic organization of the pt. lexicon. (Ontology Learning, Information Extraction).
- Evaluate the knowledge extracted defining a methodology.
- Study the specific issue of systematic polysemy in Portuguese.
- Compare our model to other models of the Portuguese language (WordNet.PT and WordNet.BR).
- Make the resource publicly available.

Extracting the Structure of the Lexicon

- Can be thought of as a reverse engineering process.



What relations?

- Hyponymy; Hyperonymy
 - Saxofone - **instrumento musical** de sopro, feito de metal, recurvo, com chaves e embocadura de palheta
 - **is_a**(saxofone, instrumento musical)
- Meronymy; Holonomy
 - rim – **orgão** que tem a a função de...
 - órgão – cada uma das **partes do corpo**...
 - **is_a**(rim, órgão) & **part_of**(órgão, body) -> **part_of**(rim, body)

What relations (cont'd)?

- **Synonymy**

- permutar – **trocar**;
 - **syn**(permutar, trocar)

- **Antonymy**

- infeliz – o que não é **feliz**
 - **ant**(infeliz, feliz)
- irracional – não **racional**
 - **ant**(irracional, racional)



Morphological processing:
infeliz = **in** + feliz
descontente = **des** + contente

What relations (cont'd)?

- Causation

- matar - causar a **morte** a
 - *causa*(matar, morte)

- Entailment

- ressonar - respirar com ruído durante o **sono**
- sono – estado de quem **dorme**
 - *entails*(ressonar, dormir)

- Cross part-of-speech relations

- informatização - acto ou efeito de **informatizar**
 - *nominalization*(informatizar, informatização)

Extracting the Structure of the Lexicon

Árvore -- planta lenhosa que pode atingir grandes alturas e cujo tronco se ramifica na parte superior

árvore (*tree*)

=> planta lenhosa (*woody plant*)

=> organismo (*organism*)

=> ser vivo (*living thing*)

=> ente (*entity*)

Structure the Lexicon

(Simple English example)

Tree -- a tall perennial woody plant having a main trunk and branches forming a distinct elevated crown; includes both gymnosperms and angiosperms.

tree

=> woody plant

=> vascular plant

=> plant

=> organism

=> living thing

=> physical object

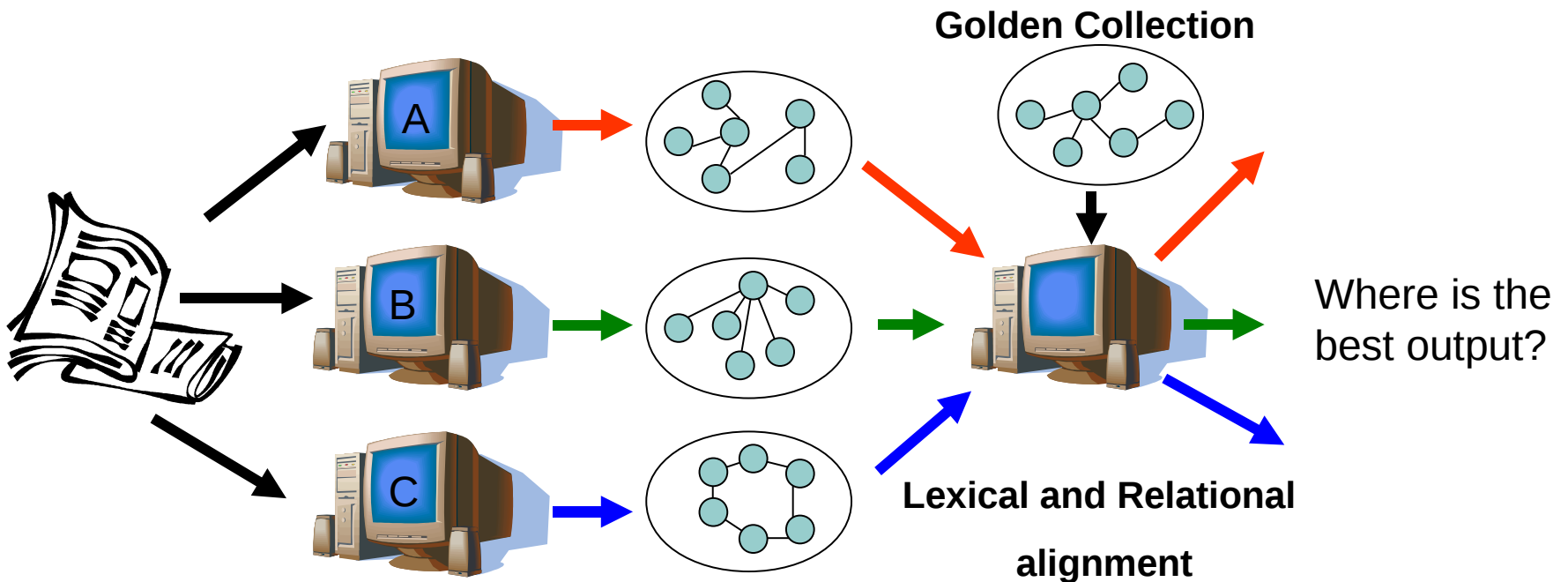
=> entity

Taken from WordNet 2.1

Ontology Evaluation

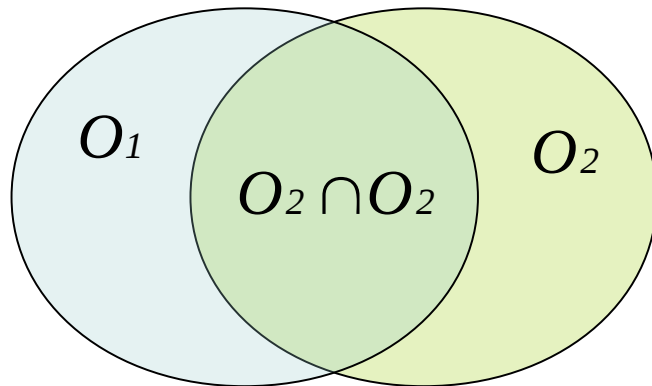
- Evaluation has received very little attention!!
- But still, we can identify 4 core kinds:
 - The use of a golden collection
 - Evaluate the output of some ontology driven process
 - Compare the ontology with clusters generated from corpora
 - Human evaluation

Using a Golden Collection



Using a Golden Collection (cont'd)

- At the lexical level (terms in common)
 - Precision, Recall, F-Measure, ...

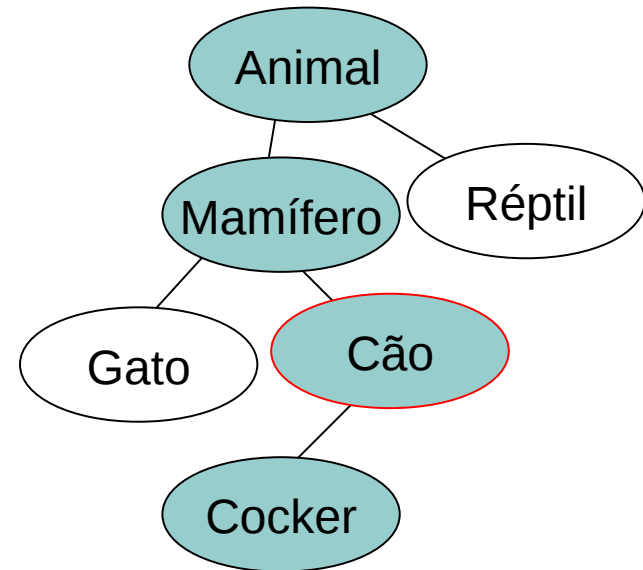
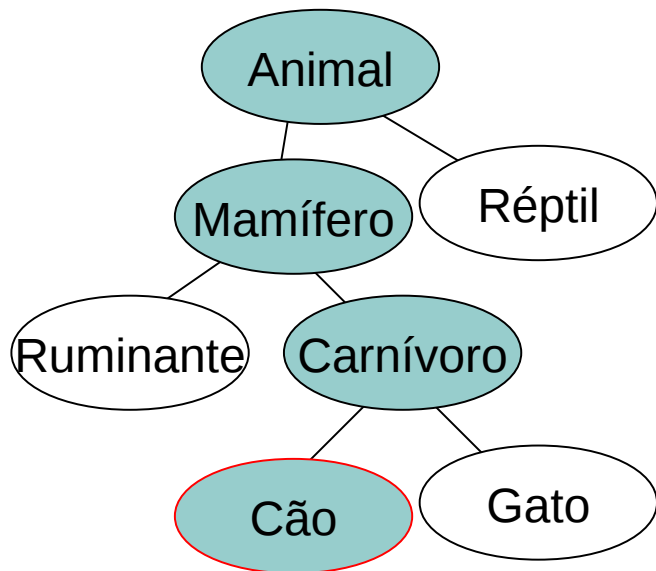


$$Pr = \frac{O_1 \cap O_2}{O_1}$$

$$Abr = \frac{O_1 \cap O_2}{O_2}$$

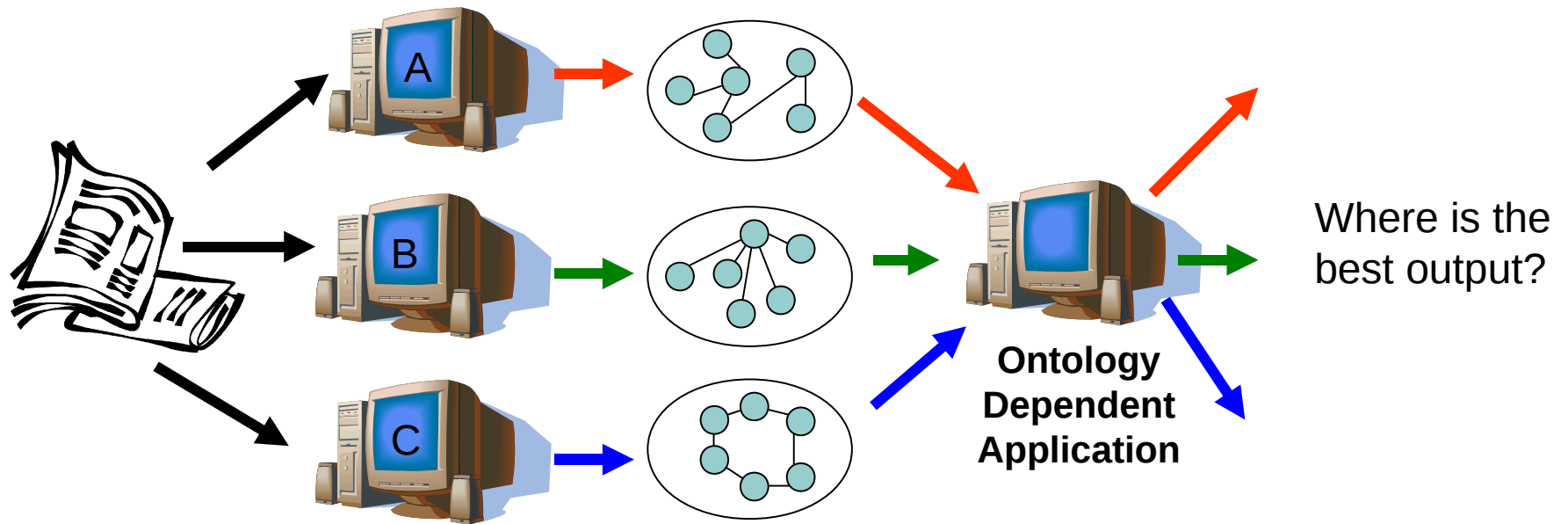
Using a Golden Collection (cont'd)

- At the relational (hyperonymy/hyponymy) level (Maedche et al., 2002)



$$TO(c\tilde{a}o, O_1, O_2) = \frac{3}{5}$$

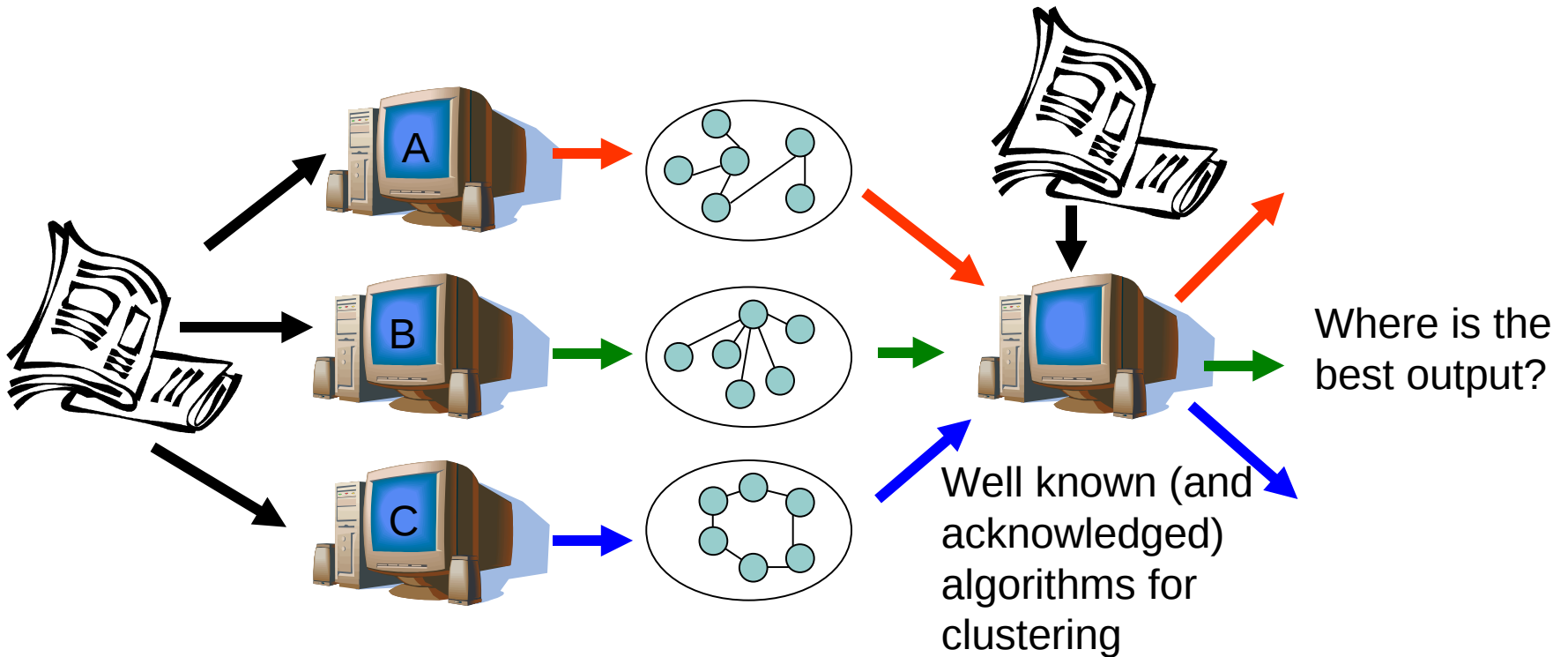
Evaluate the Output of an Ontology Dependent Application



Evaluate the Output of an Ontology Dependent Application (cont'd)

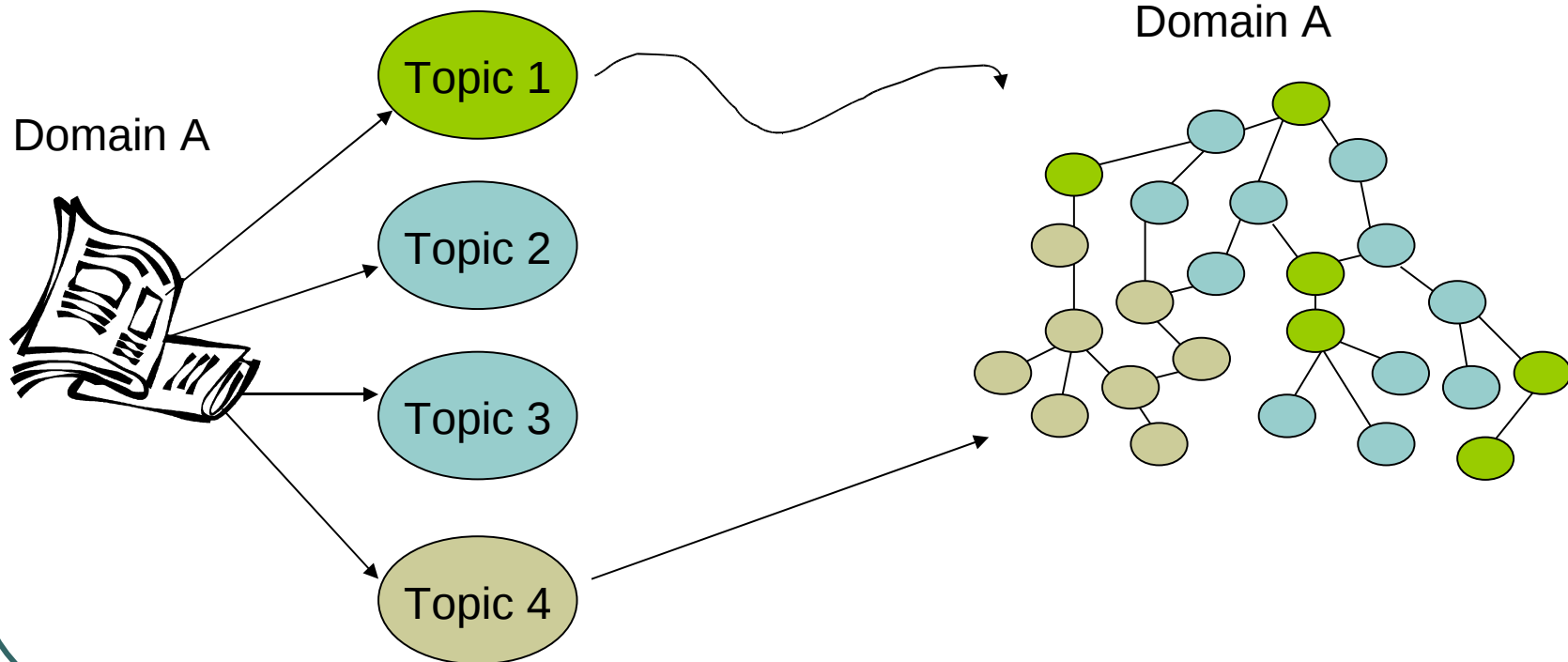
- Semantic similarity computations using ontologies and correlating them with human judgments.
- Performing query expansion in information retrieval systems.

Use clustering strategies (coarse evaluation)

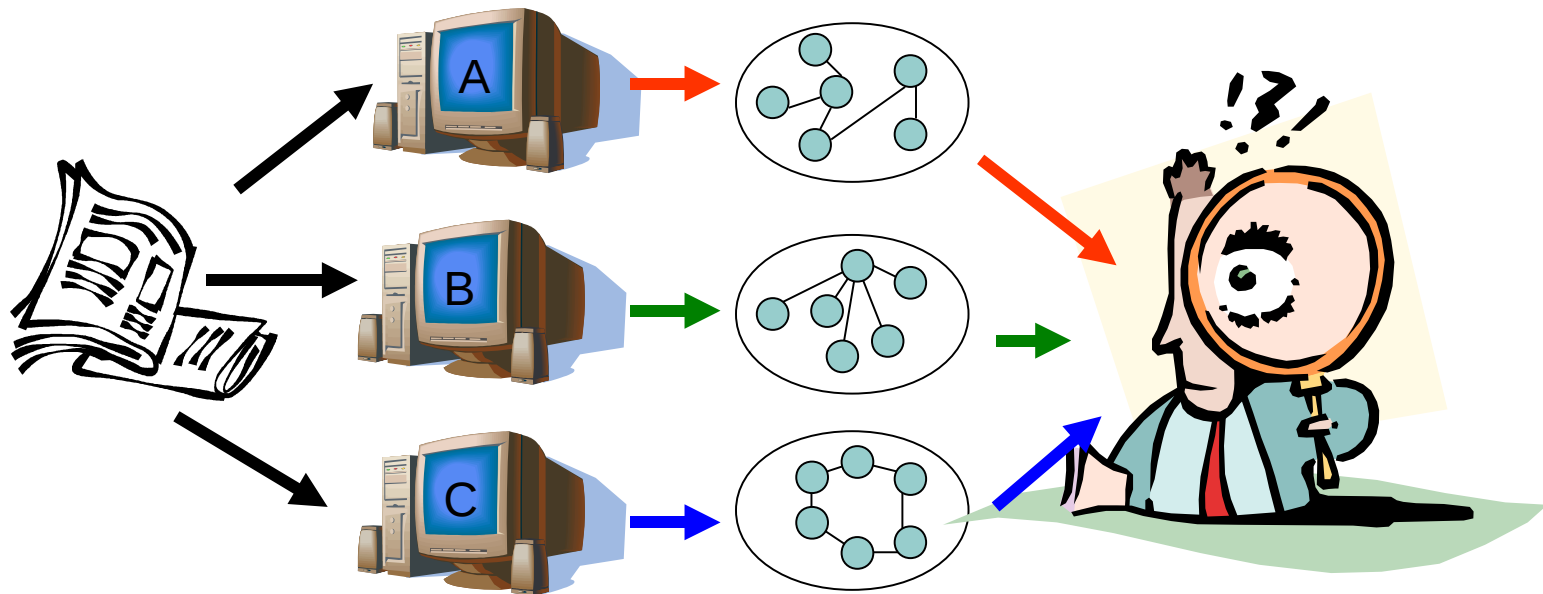


Use clustering strategies (coarse evaluation)

- Brewster et al., 2004



Human evaluation



Human Evaluation (cont'd)

- In order to ease the evaluators task, one could show the definitions for each (new) concept in the ontology. (Navigli et al.):
 - festival – “a day or period of time set aside for feasting and celebration”
 - jazz – “a style of dance music popular in the 1920s; similar to New Orleans jazz but played by large bands”
- ↓
- jazz festival – “a kind of festival, a day or period of time set aside for feasting and celebration, related to jazz, a style of dance music popular in the 1920s”

How can I evaluate my work?

- Manual Inspection !
- Compare to other resources being constructed:
 - Luís Sarmiento (Linguteca, Porto) – extracting relations from corpora.
 - Marcírio Chaves (Linguteca, Lisboa) – creating a geographical ontology.
- Feed the ontology to ongoing projects:
 - AI Lab - ReBuilder
 - Linguateca, Oslo - Esfinge .

Word senses: Polysemy vs. Homonymy

- An individual word or phrase that can be used (in different contexts) to express two or more different meanings.
 - **Polysemy** - senses are related in some way (complementary).
 - School starts at 8:30.
 - The School was founded in 1910
 - **Homonymy** - senses are unrelated (contrastive).
 - The bank has several offices.
 - We walked along the bank of the river.

Systematic Polysemy

*“Polysemy of word **A** with meanings \mathbf{a}_i and \mathbf{a}_j is regular [systematic] if there exists at least one other word **B** with meanings \mathbf{b}_i and \mathbf{b}_j which are semantically distinguished from each other in exactly the same way as \mathbf{a}_i and \mathbf{a}_j and if \mathbf{a}_i and \mathbf{b}_i , and \mathbf{a}_j and \mathbf{b}_j are nonsynonymous.”*

Ju. Apresjan (1974)

Some examples...

- Habitante/Língua (Habitant/Language)
 - norueguês, português, escocês, ... (68)
- Fabricante/Vendedor (Producer/Seller)
 - pasteleiro, ourives, queijeiro, ... (57)
- Abertura/Acto (Opening/Act)
 - vista, entrada, perfuração, ... (11)

Role of Systematic Polysemy

“Acknowledging the systematic nature of polysemy and its relationship to underspecified representations allows one to structure ontologies for semantic processing more efficiently, generating more appropriate interpretations within context”

Paul Buitelaar (1998)

Progress so far...

- Studying the physical format of the dictionary of Porto Editora, *Dicionário da Língua Portuguesa*.
- Looking for frequent patterns, indicative of interesting relations.
- Parsing the definitions using some of these patterns to obtain a taxonomic structure to the lexicon.
- Preliminary mining of systematic polysemy patterns.

Building a Large Scale Lexical Ontology for Portuguese

Nuno Seco

Linguatca Node of Coimbra

<http://linguatca.dei.uc.pt>

The Dictionary in Numbers

- Porto Editora's Dictionary (open class words)
 - Number of entries:
 - Nouns - 61980
 - Verbs - 12378
 - Adjectives - 26524
 - Adverbs - 1280
 - Number of senses:
 - Nouns - 110451
 - Verbs - 35439
 - Adjectives - 44281
 - Adverbs - 2299

The Dictionary in Numbers

- Frequent patterns in noun definitions:
 - acto ou efeito de ... (3851)
 - pessoa que ...(1386)
 - indivíduo ... (1235)
 - aquele que ... (1148)
 - parte ...(1052)
 - conjunto de ... (1004)

The Dictionary in Numbers

- Frequent patterns in verbs definitions:
 - fazer ... (1680)
 - tornar ... (1359)
 - tirar ... (744)
 - pôr ... (674)
 - causar ... (299)
 - estar ... (284)

The Dictionary in Numbers

- Frequent patterns in adjective definitions:
 - que tem ... (2698)
 - que ou aquele que ... (1393)
 - relativo a/ao/à ... (1236+725+1162)
 - relativo ou pertencente... (647)
 - que ou o que ... (527)
 - que diz respeito ... (494)

The Dictionary in Numbers

- Frequent patterns in adverb definitions:
 - de modo... (393)
 - de maneira ... (48)
 - do ponto de vista ... (28)
 - por meio de ... (14)

Some difficult issues...

- Finding the right sense of word in the definition:
 - arquibancada – banco grande cujo assento ...
 - What sense of banco?
- Circularity:
 - passagem – **transição** de um ...
 - **transição** – passagem que comporta ...

Complementary Studies

árvore (*tree*)

=> planta lenhosa (*woody plant*)

=> organismo (*organism*)

=> ser vivo (*living thing*)

=> ente (*entity*)

Extracted from pt dictionary

tree

=> woody plant

=> vascular plant

=> plant

=> organism

=> living thing

=> physical object

=> entity

Taken from WordNet 2.1