REPENTINO - A Wide-Scope Gazetteer for Entity Recognition in Portuguese

Luís Sarmento
Ana Sofia Pinto
Luís Cabral

Outline

- Small Introduction & Motivation
- 3 Assumptions underlying REPENTINO
- Collecting Examples
- The "Loose" Tree-like Structure
- Statistics about REPENTINO
- Evaluation: how to evaluate REPENTINO?
- Future Work
- Conclusions

Introduction

- NER Task: to recognize names of entities in text (and classify them according to context)
- Traditionally, NER Task involves classifying: People, Places, Organizations, Num. Ex. (MUC)
- NER task is evolving to analysis of a wider range of entities and finer-grained classification:
  - More top level classes: Events, Things, Abstractions
  - Many sub-classes for each category

Developing NER Systems

- Three possible strategies:
  1. Using Gazetteers and a set of manually encoded rules
  2. By inferring classification rules from a hand-tagged reference corpus using ML techniques
  3. Combination of the previous two
- In any case, important resources are needed:
  - Gazetteers
  - Reference corpus
- Are there such type of resources for Portuguese? Are they publicly available?

Main Motivation

- While developing our NER system we were not able to find a publicly available and comprehensive gazetteer
- New sub-goals:
  - Build a wide-scope gazetteer storing information about many different types of entities
  - Fill the gazetteer with examples of entities that are only now being currently considered in NER
  - Make it publicly available to all community, which may also contribute to its development collaboratively
  - Include instances in other languages if they are useful for NER purposes

Organizing Entities

- No generic guidelines to organize entities
- Old problem:
  - "How to organize objects of the world?"
- Should we organize the entities using a:
  - simple taxonomy? (ex: Sekine et al., 2002 / 2004)
  - tree-like structures with multiple inheritance?
  - graph-like structure?
- Should we *predefine* a certain ontology?
- Too many hard open questions:
  - We "just" wanted to build a NER system!!

Three Basic Assumptions for REPENTINO

1. Use a bottom-up approach:
   - Start from a simplified version of the ontology provided by the HAREM guidelines
   - the classification structure should reflect what is actually possible to find in corpora. Ex:
     - If we find a lot of references to luxury yachts then it might be worth to have a class / subclass to store them
   - Prefer a "loose" hierarchy instead of a very rigid and strict organization (leave it for later)
     - Not too deep or too restrictive which is difficult to justify

Three Basic Assumptions for REPENTINO

2. Classify instances according to "surface-structure":
   - Only consider the most immediate sense of the instance disregarding context-dependent interpretations. Ex:
     - Despite countries may be seen as a Organization in some contexts, they will be classified as specific types of Places
   - Ambiguous cases, such as PLACE / ORG, once decided for a particular instance should be resolved consistently. Ex:
     - If "Teatro D. Maria II" is classified as "cultural location" instead of "cultural organization", then "Teatro de São João" will also be classified as "cultural location"
   - All ontological inferences and ambiguity resolution problems are considered application / scenario dependent.

3 Basic Assumptions for REPENTINO

3. Store instances of **names** rather than instances of individual 'entities' or 'facts'
   - "America" is stored under Location and "America" (Kafka's book) is also stored under Art-Media-Com
     - the same name relates to two different types of entities so they are two valid instances to be stored in REPENTINO
     - A second book called "America" would **no**t be stored
     - The information stored in REPENTINO is restricted to:
       - "there is a book called 'America'"
     - not that there is a book by Kafka (which is another 'fact')

Building REPENTINO

- Process guided by problems found during the development of our NER system
  - When our NER system was not able to classify a given NE (ex: a luxury yacht) we would try to gather more examples of such a type of NE
  - Considerably enriched REPENTINO diversity
- Two options for finding new examples:
  - Searching large corpora using simple patterns
  - Using Google and specific domain web sites

Using Corpora and Patterns

- Using BACO, a MySQL encoded version of the WTP03 web document collection, we performed a search for archetypical patterns :
  1. Typical head-words: "Universidade d*"
  2. Archetype collocations / contexts: "localizado em *", "próximo d*"
  3. Typical endings: "Lda.", "S.A."
- Each round would take about two hours including manual validation:
  - about 1000 candidates each round

Using Google and specific domain web sites

- For many very specific entities (eg: "computer games", "music", etc), it became very difficult to find efficient search patterns
- But there are many web sites specifically dealing with such domains
  - Searching Google with those keywords or other search expression, such as "list of celebrities", it became very easy to find huge lists of instances, ready to be stored
- Other interesting sites:
  - Wikipedia, stock exchange sites, "warez" sites

The “loose” hierarchy

- Currently REPENTINO is organized in
  - 11 top categories
  - 97 subcategories
- Weak ontological relationships between the top categories and corresponding subcategories
  - REPENTINO could be simply seen as a collection of 97 specialized gazetteers that may be re-organized according to a specific application scenario

The “loose” hierarchy (1)

Location Entities that are individualized essentially according to their position in the Universe.	Terrestrial, Hydrographic, Address, Loose Address, Country/State, Town/Region/Administrative Division, Space, Socio-Cultural, Religious, Civil/Administration/Military, Heritage/Monuments, Other, Real-Estate, Mythological/Fictional, Commercial/Industrial/Financial, Infrastructure/ Facility.
Organizations Entities that are composed by more than one person and that exist and operate as a whole. Organizations usually have goals, a set of rules and an internal structure that rule them, as opposed to simple groups of people or gatherings.	Company, Government/Administration, Education/R&D, Sports, Socio-Cultural, Interest Groups, Religious, Civil/Military, Clubs, Section, Other. Beings
Beings Real or fictional beings, as well as myths and mythological beings	Human, Human-Collective, Non- Human, Geopolitical/Ethnic/Ideological, Mythological, Other

The “loose” hierarchy (2)

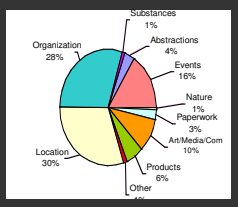
Events Events whose beginning and time span are clearly defined.	Ephemerid, Cyclic, Scientific, Socio-Cultural, Sports, Political, Prize/Award, Other.
Products This category includes many possible entities, ranging from industrial products to handcrafted objects. There is an important difference between a Product and an Organization, since a Product should refer to a specific model, while organization is its producer.	Brands, Consumables, Electronics/Appliances, Financial, Format, Gastronomic, Inspection/Exam, Services and Resources, Computer Systems and Applications, Clothing/Utilities, Vehicles, Medical/Pharmaceutical, Tools/Instruments, Craftwork, Other.
Art/Media/Communication This is a specialized category that deals uniquely with products related to art, media and communication	Books, Movies, TV/Radio/Theatre, Music, Fine-Arts & Design, Multimedia, Periodical, Scientific/Academic Paper, Other.

The “loose” hierarchy (3)

Paperwork Laws, Decrees, Treaties, Pacts, Standards, Rules, Documents, Taxes and alike should be included in this category	Laws, Certificates, Documents, Taxes/Fees, Proof/Test/Evaluation, Agreements, Standards, Other.
Substance In this category we include elements, substances and minerals.	Group, Ore, Substance, Other.
Abstraction Abstract entities such as disciplines, sciences, crafts, as well as certain mental formulations. We also include specific time periods, movements, states, diseases and crimes.	Disciplines/ Crafts, Period/Movement/Trend, State or Condition, Mental Formulation, Symbols, Crime, Latin Expressions, Era, Process, Type/Class, Index/Tax, Other.
Nature This category includes animals, vegetables, all the elements that constitute living beings, as well as natural phenomena.	Animal, Physiology, Micro-organisms, Vegetable, and Natural Phenomena.
Miscellanea In this category we include words or symbols that are susceptible to collocate or to be present in the near context of some of the previous entities.	Personal titles, Currency Units, Others.

Current Figures

- Number of instances: 450.000
- Larger Top-Category: Beings - 288.000 - 65%
- Distribution of remaining 35%:



Evaluation

- We have not yet performed any **direct** evaluation
  - Is there a way of evaluating gazetteers?
- What can be directly measured?
  - Size? Variety?
  - How many capitalized words from a large corpus can be found in REPENTINO?
  - How many entities from a “golden collection” exist in REPENTINO?
- But what do these values mean?
- How to compare with other repositories:
  - Should we compare their contents?
  - Should we compare their underlying ontologies?

Evaluation

- Indirect Evaluation: does the gazetteer help in a specific application scenario.
  - Obvious task: NER
- We used REPENTINO as the main gazetteer in SIEMÊS:
  - Good global results in HAREM but…
  - we are still waiting for the results of the component evaluation of SIEMÊS (in mini-HAREM this April) to see how much did REPENTINO actually help (or not!)

Future: How to improve REPENTINO?

- Obtain more information about each instance, using the Web:
  - Frequency counts for each instance
  - Co-occurrence counts (info about NE clusters?)
  - Contexts on a sub-class basis. For example: what are the most common contexts for "Locations"
- Obtain more instances, ideally very different from the ones already stored:
  - Users can suggest examples through the web site

How to get REPENTINO

- http://www.linguateca.pt/repentino/
  - search interface
  - Interface for submitting new instances
  - XML file containing the entire data set
- http://www.fe.up.pt/~las/ (see "software")
  - A Perl module containing ready to use information and function