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

Luís Sarmento, Ana Sofia Pinto, and Luís Cabral

Faculdade de Engenharia da Universidade do Porto (NIAD&R),
Rua Dr. Roberto Frias, s/n 4200-465 Porto, Portugal
las@fe.up.pt
Linguateca – Pólo do Porto, Portugal Via Panorâmica s/n, 4150-564 Porto
asofia@letras.up.pt
Linguateca – Pólo de Oslo,
P.O. BOX 124, Blindern, 0314 Oslo, Norway
Luis.M.Cabral@sintef.no

Abstract. In this paper we describe REPENTINO, a publicly available gazetteer intended to help the development of named entity recognition systems for Portuguese. REPENTINO wishes to minimize the problems developers face due to the limited availability of this type of lexical-semantic resources for Portuguese. The data stored in REPENTINO was mostly extracted from corpora and from the web using simple semi-automated methods. Currently, REPENTINO stores nearly 450k instances of named entities divided in more than 100 categories and subcategories covering a much wider set of domains than those usually included in traditional gazetteers. We will present some figures regarding the current content of the gazetteer and describe future work regarding the evaluation of this resource and its enrichment with additional information.

1 Introduction

The importance of Named Entity Recognition (NER) systems has been growing with the widespread of information extraction systems and applications. The goal of NER is to identify and correctly classify all Named Entities that exist in a given text according to a given predefined hierarchy or ontology. Broadly speaking, Named Entities (NE) include all entities that may be identified by a proper name, such as, for example, people, organizations, places, brands or products and other more abstract classes such as knowledge domains, techniques, or intellectual products (e.g.: “Computational Linguistics” or “9th Symphony”). The classification of numeric and time references is also usually included in the NER task. However, the detail and complexity of this task has varied greatly and has evolved over time. For example, in the first NER evaluation programs during MUC-6 [1], systems were asked to identify and classify entities belonging to a small set of generic categories, namely Person, Organization or Location. More recent evaluation programs, such as the ACE [2] or the Portuguese evaluation effort HAREM [3], required systems to perform classification over more detailed two-level hierarchies, to determine the semantic role of the referenced entities and to deal with other complex contextual constructions [4].

R. Vieira et al. (Eds.): PROPOR 2006, LNAI 3960, pp. 31 – 40, 2006.

Most NER systems are built using two possible strategies: (i) gazetteers and a set of manually encoded rules or (ii) by inferring classification rules from previously annotated corpora using supervised machine learning (ML) methods. In both cases important language resources are required (i.e. gazetteers or annotated corpora). Unfortunately, some languages lack publicly available resources for these purposes and adapting existing resources from other languages may require an effort equivalent to that of building the resource from scratch. Portuguese is one of those languages where the lack of resources has been problematic for the development of NER systems. Therefore, developing such resources should be considered a strategic option for the research community working with the computational processing of Portuguese. Additionally, since the definition of the NER task is rapidly expanding to include many more categories than the traditional ones (organization, location, person and numeric expressions), existing resources, when available, may not be enough to cover these latest requirements, which demand wider-scope gazetteers.

In this paper we will present REPENTINO (REPositório para reconhecimento de ENTIdades com NOme), a new publicly available gazetteer we have been developing that is organized according to a wide hierarchy that includes 11 top categories and 97 subcategories. Currently, REPENTINO stores more than 450000 instances of NE that have been extracted mainly from a large document collection and also from several thematic Web sites. REPENTINO has been manually validated to ensure the quality of the resource and is now freely available online in XML format from http://www.linguateca.pt/repentino/.

2 Motivation

Our motivation for building REPENTINO came from the difficulties encountered during the development of our own NER system: we were not able to find appropriate gazetteers for our NER system, either because they did not cover all the categories we were considering or because, for some categories, they were not comprehensive enough, covering only a small fraction of the cases. We thus began studying the possibility of building our own gazetteer, by making use of simple extraction techniques. The kind of techniques we were considering consisted of searching large quantities of text for archetypical lexical patterns that could lead us to instances of named-entities. For example, the lexical pattern “located in [Uppercased string]” could be used to identify instances of geographical entities. Although this approach seems quite naïve at first, simple tests allowed us to confirm that it is actually possible to extract hundreds of instances of organizations, locations and events from corpora with such simple techniques. Most importantly, instances could be validated without too much manual effort. Such procedures have, of course, their own limitations: it is very difficult to extract instances of some classes, such as product brands (e.g.: “luxury yachts”) or companies, because the contexts in which they appear are more diverse and more difficult to identify. But on the other hand, there are innumerous web sites where it is possible to find long lists of instances of such NE, and in some cases it is quite easy to harvest them manually by simple “copy-and-paste” methods. These two possibilities seemed promising enough to invest some effort in building a wide scope database of manually classified NE instances, for NER purposes.

3 Related Work

The need for wide scope classification systems capable of dealing simultaneously with various Information Extraction scenarios has been pointed out by Sekine [5]. The authors present an extended NE classification hierarchy, which includes 150 different types of NE organized in a 3 level tree structure, aiming to be comprehensive enough to cover major newspaper domains. This hierarchy is intended to classify a wide range of possible NE, including even some entities that may occur without capitalization (e.g. “leukemia”). In a later work [6], this hierarchy was extended to 200 categories. The authors also developed a dictionary containing 130000 instances of named-entities, organized according to the hierarchy developed. Instances were manually compiled from the Web, newspaper and other sources. The hierarchy was populated considering only the surface form of the entities. For example, “Portugal” would be considered a Place, although it may adopt different senses depending on the context (e.g.: “Organization”). In order to deal with several frequent cases of ambiguity in NE classification (e.g.: museums, theatres as either “Places” or “Organizations”), the hierarchy has several diffuse categories intended to classify such ambiguous instances.

Other recent works focus on dynamically building open classification hierarchies. Pasca [7] describes a system that is capable of finding both the NE instances and the corresponding (multiple) categories, using a lightly supervised Machine Learning (ML) algorithm. The author argues that traditional approaches to NE classification face the strong limitation of using closed categorization hierarchies, which most of the times are too coarse for dealing with flexible information extraction scenarios, namely web search. In those cases, categories are very diverse, overlapping, and very specific, which makes the process of developing a pre-defined category and the corresponding gazetteer unfeasible. Starting from a set of domain independent extraction patterns, the system is able to find categorized instances of named entities, and to obtain new contextual extraction patterns to find more categorized instances. The system is able to infer both general and very specific categories (and to obtain the corresponding instances) such as “colors”, “hybrid cars” or “operating systems”.

Our work for developing the classification system of REPENTINO and acquiring the corresponding NE instances lies somewhere between the top-down strategy followed by Sekine and the bottom-up approach of Pasca’s work. Because of this, during the development of REPENTINO’s hierarchy, we faced similar problems to those described in [5] such as for example deciding if a given instance should imply the creation of a new class in the system or could it be easily fit in an existing one. At the same time, our strategy for compiling instances of NE to populate REPENTINO has some points in common with the techniques described in [7] - although using manual processes instead of ML techniques – and has lead us to include several categories in the classification structure that we would have never otherwise predicted.

4 Structuring REPENTINO

The most complex questions when developing lexical-semantic resources are related to data organization. In developing a wide scope, fine-grained resource those questions involve dealing with philosophical assumptions about how the world should be organized. We knew of very detailed NE hierarchies, like [5] and [6], but such fine-grained hierarchies are very rare and, to our knowledge, there are no generic guidelines available for building them. There are many difficult questions related to developing adequate classification structures. For instance, in hierarchical structures, deciding if a given category should be split in several more specific may not be trivial and usually leads to some arbitrary decision. One should also note that, because of the ambiguous nature of many entities, a simple taxonomic hierarchy may not be an adequate structure at all, and may lead to difficult decisions regarding the classification of certain instances that could easily be placed in more than one category. Multiple inheritance connections may help to solve some of these questions but this usually leads to more complex classification systems. In fact, the whole issue of developing a classification structure is even more complex than that since any classification structure implies a specific ontology. However, any ontology (when agreed upon) is usually application-dependent, so committing to a given ontology may reduce the generality and portability of the resource.

Therefore, for building REPENTINO we followed three basic assumptions. The first assumption is that the classification structure should reflect the instances actually found in corpora or on the web, and should not be a pre-defined closed hierarchy, which usually involves making several ontological commitments. We thus decided not to adopt a pre-defined closed hierarchy but, instead, to follow an ad-hoc strategy for expanding an open set of categories, based on the instances that we were able to actually collect from corpora and from the Web by the processes described in the next sections. The structure of REPENTINO may be seen as a “loose” two-level hierarchy, with several generic top categories where more specialized sub-categories are spawned as new interesting NE instances are found. The hierarchy is not based on any strong ontological relations between top-level categories and their sub-categories. We tried to remove as many ontological assumptions as possible from the classification structure to make REPENTINO’s content reusable in several scenarios, and to circumvent hard philosophical questions. Sub-categories are considered the core of REPENTINO: instances are directly connected to subcategories and top-level categories which exist mainly for convenience reasons. The sub-categories could exist independently of the top-level categories, for example, as several separate specialized gazetteers. Ontological relations among instances or sub-classes are outside the scope of REPENTINO, and, if needed, they should be implemented by a particular application.

The second assumption is that instances found would always be classified according to their surface structure, i.e. considering the most immediate sense of the instance, and totally disregarding the several possible senses in context. Ambiguous cases, such as the place / organization ambiguity, once decided for a particular instance (e.g.: “Teatro Nacional D. Maria II” as a “cultural place” and not as an “organization”), would automatically imply that all similar cases would be classified equally (e.g. “Teatro de São João” would also be classified as a “cultural place”). For example, countries are stored in REPENTINO as a specific type of place. Ontological inferences, such as “a country may be seen as an organization in certain contexts”, are not in any way implied in REPENTINO, and depend solely on the application that uses the information stored in REPENTINO.

The third assumption is that REPENTINO stores instances of names rather than instances of individual entities or facts. This is indirectly related to how homograph instances should be dealt with. For example “America” may refer (at least) to a continent or to Franz Kafka’s book. Obviously, these two instances should be stored separately in REPENTINO, under the two different corresponding subcategories (in this cases Location-Terrestrial and Art/Media/Communication-Book as it will become clear later). But let us assume that there is another book named “America”. Should we store a second entry “America” under the same subcategory we place Kafka’s book before? The answer is negative because REPENTINO is intended to store names, not facts. REPENTINO should simply provide the information that there is (at least) one book named “America” (or that “America” could refer to an existing book) but not that “America” is a book by Franz Kafka, or by any other author.

5 Building REPENTINO

The actual process developing REPENTINO was very dynamic and was guided by particular problems faced during the development of our NER system. Whenever a given entity could not be correctly classified by our NER system - for example a luxury yacht brand - rather than trying to create a rule to deal with this case, we would search corpora or the Web for more instances of similar entities. This allowed us to obtain a broader picture of the problem and good insights about whether those instances should be added to REPENTINO or not.

This strategy affected dramatically the development of REPENTINO’s hierarchy. For instance, we were thus lead to create 16 subcategories under the category Location, almost all of which with more than 100 instances, and some with more than a thousand instances. But more importantly, we were able to include some very frequently mentioned named entities - such as Infrastructure/Facility or Real-Estate - that are rarely considered in most NE classification hierarchies. Similar situations happened for other top categories. It was also possible to compile many other instances that allowed us to include in REPENTINO totally unorthodox categories. For instance, REPENTINO includes a top category named “Paperwork” which we were able to fill with about 4500 instances, divided into eight subcategories.

5.1 Collecting NE Using Simple Patterns and Corpora

For extracting instances of NE from free text, we used BACO [8], a text database generated from the 14Gb WPT03 collection (http://www.linguateca.pt/wpt03/). The WPT03 collection is quite recent so it is very appropriate for extracting instances of relevant NE. However, for the extraction process to be feasible, we needed to be able not only to identify lists of possible NE instances, but also to have very strong clues about their categories to reduce the effort of manual validation. We thus tried to explore morphological and contextual properties of the (Portuguese) NE:

1. a typical head-word. Most of the entities have typical head-words, i.e. the first words of the NE are very typical and can almost certainly define its category: “Universidade do Porto”, or “Junta de Freguesia de Ramalde”.
2. an archetype context or collocation. There are many archetype contexts or collocations that may be used to extract certain types of NE. For example, in looking for locations, we may try to find what matches certain patterns such as “localizado na XXX” (“located in XXX”) or “próximo da XXX” (“near XXX”), where XXX has a very high probability of being a location.
3. a typical end-word. Some entities, such as companies and other organizations, may have certain typical end-words or acronyms. For example, in Portuguese, company names frequently end with particles such as “Lda.” or “S.A”.

Searches were performed using simple Perl scripts. Each complete run took approximately two hours, including manual revision, and we were usually able to extract up to 1000 instances of NE per run (sometimes many more).

5.2 Retrieving Instances from the Web

For some specific NE categories we found that it was much easier to find domain specific sites and collect some of the published information. For example, there are many sites on the web containing huge lists of movies, music and software applications. Such information is very difficult to extract from corpora, especially because it is not frequent enough, but it is readily available in some web sites. We were able to retrieve information from over 120 websites, taking advantage of several thematic ones, which did not have to necessarily be Portuguese. For example, names of software products, movie stars from the sixties or of luxury yachts can be compiled from sites in many possible languages. Apart from large scope sites such as the Portuguese and English version of Wikipedia, a great deal of our collecting effort was done over domain specific sites as, for example, sites from stock exchange markets. The choice of these domain specific sites was done in a rather ad hoc way. Some of them were found after searching the web using a regular search engine with a set of seed entities or by explicitly entering search expressions such as “list of celebrities”.

Other resourceful sites were found by trying links from well-known institutional sites. For example, we were able to find lists of several pharmaceutical products and active chemical substances visiting the web site of the national pharmaceutical administration office. Despite the apparent randomness of the process that led to many dead ends, this strategy proved to be an appropriate technique for collecting instances of NE that could not be easily retrieved from corpora. We believe that this allowed us to greatly improve the diversity of REPENTINO.

6 The “Loose” Classification Hierarchy of REPENTINO

Presently, the REPENTINO hierarchy comprises 11 top categories and 97 subcategories. Note that many of the subcategories are not likely to be considered when building a hierarchy using a top-down approach. However, by the processes explained before, we were able to retrieve large quantities of instances for such categories, which justifies their inclusion in REPENTINO. We will now present the current categories and subcategories and provide a brief explanation about them.

Location
Entities that are individualized essentially according to their position in the Universe. This category comprises the following subcategories: 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. Organizations are divided in the following subcategories: Company, Government/Administration, Education/R&D, Sports, Socio-Cultural, Interest Groups, Religious, Civil/Military, Clubs, Section, Other.

Beings
Real or fictional beings, as well as myths and mythological beings. Additionally, groups of people that do not explicitly form an Organization, such as ethnic and geopolitical groups, are also part of this category. In this hierarchy, the difference between Fictional beings and Myths is mainly that Fictional characters have never existed while Myths are not guaranteed to have existed or not. Also, a separate subcategory is considered for mythological beings, which are not the same as Myths. Beings are divided in the following subcategories: Human, Human-Collective, Non-Human, Geopolitical/Ethnic/Ideological, Mythological, Other.

Event
Events whose beginning and time span are clearly defined. Events include the following subcategories: Ephemerid, Cyclic, Scientific, Socio-Cultural, Sports, Political, Prize/Award, Other.

Products
This category includes many possible entities, ranging from industrial products to handcrafted objects. Note that although products and organizations may have a very similar name, 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. Products can be divided in the following subcategories: 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. Art/Media/Communication comprises the following subcategories: Books, Movies, TV/Radio/Theatre, Music, Fine-Arts & Design, Multimedia, Periodical, Scientific/Academic Paper, Other.

Paperwork
Laws, Decrees, Treaties, Pacts, Standards, Rules, Documents, Taxes and alike should be included in this category. This category can be divided in eight subcategories:

Laws, Certificates, Documents, Taxes/Fees, Proof/Test/Evaluation, Agreements, Standards, Other.

Substance
In this category we include elements, substances and minerals. Substances can be divided in the following subcategories: 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. Abstractions can be divided into the following subcategories: 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. Nature can be divided in five subcategories: 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 such as personal titles, currency and units.

7 Current Figures Behind REPENTINO

REPENTINO stores nearly 450000 instances of NE (a complete and updated statistical analysis is available on the web site). Currently, around 288K of the instances stored in REPENTINO (about 65%) belong to the category Beings. The distribution of the remaining instances is given in the next chart:

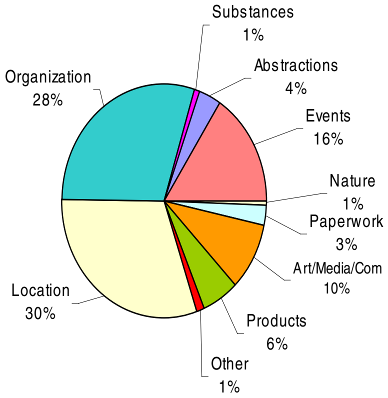

Fig. 1. – The distribution of instances according to the top categories

Apart from the category Beings, most of the instances stored in REPENTINO are Locations, Events, and Organizations, which seem to occur very frequently in the

WPT03 document collection. Other categories, such as Products are more difficult to obtain because they do not follow a strict morphology and, therefore, could not be so easily found by pattern matching processes.

8 Evaluation and Future Work

We have not yet performed any specific evaluation of REPENTINO, so no direct evaluation data is available at this moment. Direct evaluation of REPENTINO seems rather difficult because the value of this resource should be seen in relation to the success in Information Extraction tasks for which it was originally developed. At this level, some good indications about REPENTINO may be obtained by examining the results of the SIEMÊS [9], our NER system, in the recent HAREM evaluation contest. SIEMÊS heavily relied on REPENTINO as its main gazetteer and since it was one of the top scoring systems we may assume that some of its success is due to REPENTINO. A more direct evaluation of REPENTINO would have to focus on measuring specific values, such as for example the amount of overlap between its content and a gold standard, a corpus or other similar gazetteers. This will be object of future work. Other future improvements in REPENTINO aim at expanding the information stored in REPENTINO for NER purposes. For example by using a large document collection, or the Web, we may obtain information about the number of occurrences of each instance in REPENTINO and to retrieve corresponding contexts that may be used for developing rules in future NER classification procedures. Additionally, and following some of the ideas reported in [10], it seems useful to obtain information about which instances co-occur and from there try to determine possible NE clusters. Such information could be helpful for implementing new NE disambiguation procedures.

9 Conclusions

In this paper we have presented REPENTINO, a novel publicly available resource that may help researchers in the development of NER systems for Portuguese. REPENTINO was built using simple and semi-automatic NE extraction methods over large document collections, and also by manually searching the web. REPENTINO stores approximately 450000 manually validated instances of NE, organized in a loose two-level hierarchy with 11 top categories and 97 subcategories. REPENTINO has already been used in a practical NER system, whose performance was tested in the recent HAREM evaluation contest with positive results, so we believe it can be of great interest to the community developing technology for Portuguese.

Acknowledgements

This work was partially supported by grants POSI/PLP/43931/2001 and SFRH/BD/23590/2005 from Fundação para a Ciência e Tecnologia (Portugal), co-financed by POSI. The authors wish to thank Débora Oliveira for her help in the compilation process.

References

1. Grishman, R., Sundheim, B.: Message Understanding Conference - 6: A Brief History In Proc. Int. Conf. on Computational Linguistics, Copenhagen (1996) pp. 466-471.
2. Doddington, G., Mitchell A., Przybocki, M., Ramshaw, l., Strassel, S, Weischedel, R.: The Automatic Content Extraction (ACE) Program: Tasks, Data, and Evaluation. In: Proc. 4th Int. Conf. on Language Resources and Evaluation, Lisboa (2004) pp. 837-840.
3. Santos D., Seco N., Cardoso N., Vilela R.: HAREM: An Advanced NER Evaluation Contest for Portuguese. In: Proc. 5th Int. Conf. on Language Resources and Evaluation, Genoa, Italy (2006).
4. NIST. 2004. EDT Guidelines for English V4.2.6. http://www.ldc.upenn.edu/Projects/ACE/docs/EnglishEDTV4-2-6.PDF
5. Sekine, S., Sudo K., Nobata, C.: Extended Named Entity Hierarchy. In: Proc. 3rd Int. Conf. on Language Resources and Evaluation, Las Palmas, Canary Islands, Spain (2002).
6. Sekine, S., Nobata C.: Definition, dictionaries and tagger for Extended Named Entity Hierarchy. In: Proc. 4th Int. Conf. on Language Resources and Evaluation, Lisboa, Portugal, (2004) pp. 1977-1980.
7. Pasca, M.: Acquisition of categorized named entities for web search. In: Proc. 13th ACM Conf. on Information & Knowledge management. Washington, D.C., USA (2004) 137-145.
8. Sarmento, L.: BACO – A large database of text and co-occurrences. In: Proc. 5th Int. Conf, on Language Resources and Evaluation, Genoa, Italy (2006).
9. Sarmento, L: SIEMÊS – a Named-Entity Recognizer for Portuguese Relying on Similarity Rules. In: Proc. PROPOR 2006 - Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada. Itatiaia, RJ Brasil (2006).
10. Hasegawa, T, Sekine S., Grishman R.: Discovering Relations among Named Entities from Large Corpora. In: Proc. Annual Meeting of Association of Computational Linguistics (ACL 04). Barcelona, Spain (2004) pp. 415-422.