Rembrandt - a named-entity recognition framework

Nuno Cardoso

University of Lisbon, Faculty of Sciences, LaSIGE
ncardoso@xldb.di.fc.ul.pt

Abstract

Rembrandt is a named entity recognition system specially crafted to annotate documents by classifying named entities and ground them into unique identifiers. Rembrandt played an important role within our research over geographic IR, thus evolving into a more capable framework where documents can be annotated, manually curated and indexed. The goal of this paper is to present Rembrandt's simple but powerful annotation framework to the NLP community.

Keywords: Named entity recognition, annotation framework, entity grounding

1. Introduction

REMBRANDT is a named entity recognition (NER) system developed by the author, specially crafted to annotate documents by classifying named entities (NEs) and ground them into unique identifiers such as Wikipedia/DBpedia URLs (Cardoso, 2008). REMBRANDT was developed within the scope of the GREASE project, aiming to add geographic reasoning to search engines (Silva et al., 2006).

GREASE required a specific tool to annotate explicit and implicit geographic content from documents, such as cities, rivers or postal codes, so they can be associated to the certain geographic area of interest for each document. As many placenames can be used on different contexts (to designate persons, organizations or even different places), REMBRANDT's goal is to recognize all kinds of NEs so they can be properlydisambiguated.

REMBRANDT is also aware that some NEs are vague and/or ambiguous to the point that even humans cannot agree on a single semantic classification, so it allows NEs to have more than one classification to denote such vagueness and/or ambiguity that prevents a complete disambiguation.

The REMBRANDT NER tool participated in 2008 on the Second HAREM, an evaluation contest for NER systems in Portuguese (Santos et al., 2008), it obtained an F-measure of 0.567 for the full NER task and ranked as the 2nd best system out of 10, and ranking first out of 8 systems for the `PLACE` only scenario with an F-measure of 0.625, showing good capabilities for detecting and disambiguating geographic places. REMBRANDT can also annotate English texts, although the performance of REMBRANDT for English is not yet determined.

REMBRANDT played a role on the information retrieval (IR) systems that participated on GeoCLEF, a geographic information retrieval evaluation task (Cardoso et al., 2008), and in other more semantic-flavored retrieval evaluation tasks such as GikiCLEF (Cardoso et al., 2010) and NTCIR (Cardoso and Silva, 2010a). Within this evaluation environments, REMBRANDT evolved into a NER framework that also stores and indexes annotated documents, organises NEs and its grounded information in a relational database, and allows manual curation of its annotation and grounding information.

With a document collection fully annotated and curated with REMBRANDT's framework, one can query the database to extract information for a given NE (for instance, what are the most frequent NEs that appear on the same sentence with "Neil Armstrong [person]"?), and documents can be retrieved using semantic-based indexes in its document ranking procedure (as in "give me documents that contain references to Washington as a person, not the city").

The framework also provides a web interface where annotators can collaboratively curate annotated documents and all its extracted data, amending annotations for each document or for the whole collection.

The REMBRANDT is open-source code and avaliable at `http://xldb.di.fc.ul.pt/Rembrandt/`.

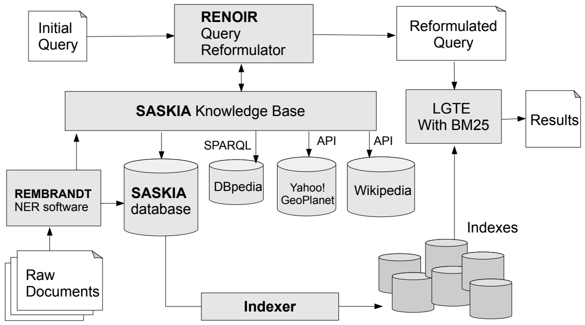
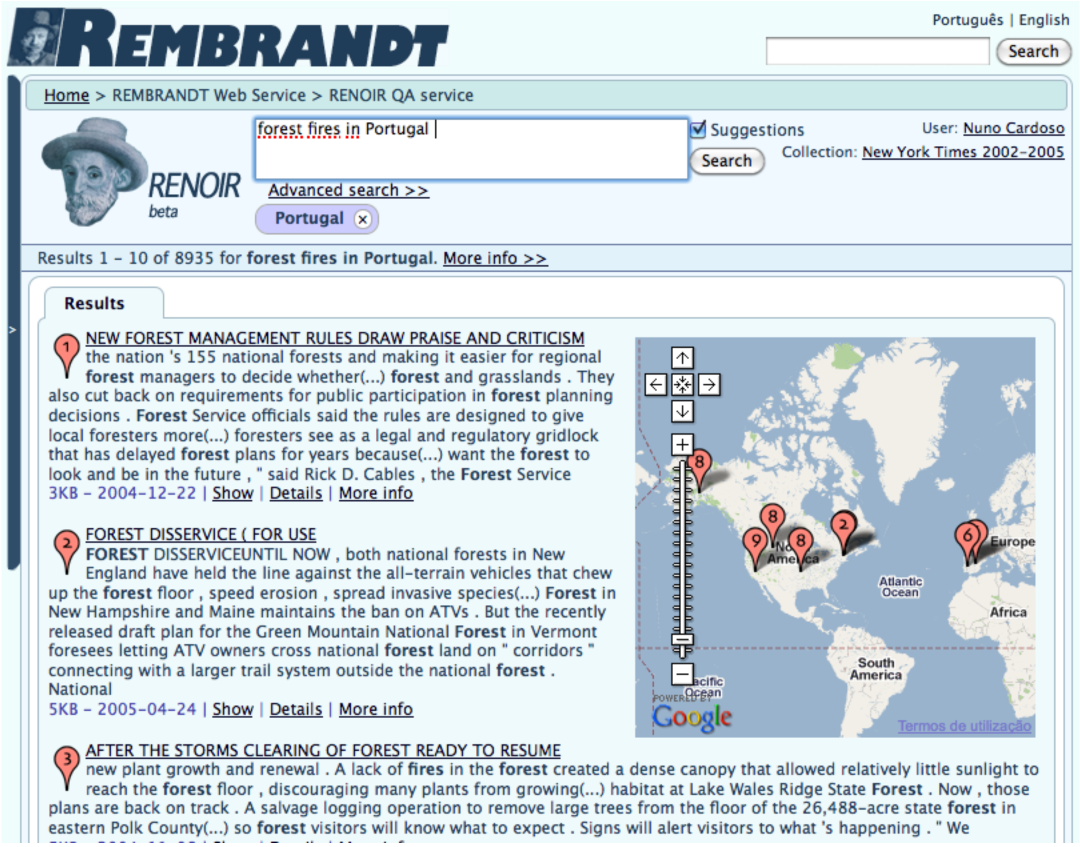


Figure 1: System overview



Português | English

Search

Home > REMBRANDT Web Service > RENOIR QA service

forest fires in Portugal

☑ Suggestions

Search

User: Nuno Cardoso

Collection: New York Times 2002-2005

RENOIR beta

Advanced search >>

Portugal ⊗

Results 1 - 10 of 8935 for **forest fires in Portugal**. More info >>

Results

1 NEW FOREST MANAGEMENT RULES DRAW PRAISE AND CRITICISM
the nation 's 155 national forests and making it easier for regional **forest** managers to decide whether(...) **forest** and grasslands . They also cut back on requirements for public participation in **forest** planning decisions . **Forest** Service officials said the rules are designed to give local foresters more(...) foresters see as a legal and regulatory gridlock that has delayed **forest** plans for years because(...) want the **forest** to look and be in the future , " said Rick D. Cables , the **Forest** Service
3KB - 2004-12-22 | Show | Details | More info

2 FOREST DISSERVICE (FOR USE
FOREST DISSERVICEUNTIL NOW , both national forests in New England have held the line against the all-terrain vehicles that chew up the **forest** floor , speed erosion , spread invasive species(...) **Forest** in New Hampshire and Maine maintains the ban on ATVs . But the recently released draft plan for the Green Mountain National **Forest** in Vermont foresees letting ATV owners cross national **forest** land on " corridors " connecting with a larger trail system outside the national **forest** . National
5KB - 2005-04-24 | Show | Details | More info

3 AFTER THE STORMS CLEARING OF FOREST READY TO RESUME
new plant growth and renewal . A lack of **fires** in the **forest** created a dense canopy that allowed relatively little sunlight to reach the **forest** floor , discouraging many plants from growing(...) habitat at Lake Wales Ridge State **Forest** . Now , those plans are back on track . A salvage logging operation to remove large trees from the floor of the 26,488-acre state **forest** in eastern Polk County(...) so **forest** visitors will know what to expect . Signs will alert visitors to what 's happening . " We

Figure 2: RENOIR user interface

2. The Rembrandt framework

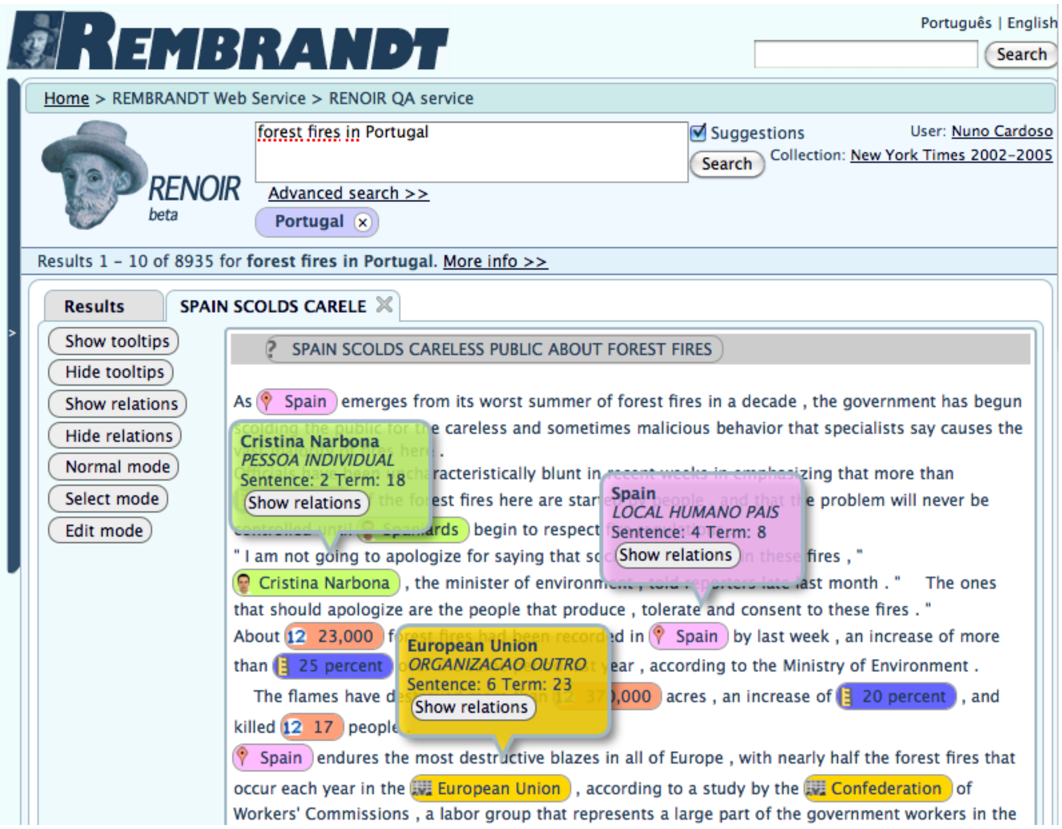
Figure 1 presents the architecture of the REMBRANDT framework, which is also described with more detail elsewhere (Cardoso and Silva, 2010b).In a nutshell, the framework is composed of i) the REMBRANDT NER tool that recognizes and grounds all entities from documents, ii) RENOIR, a semantic query reformulation module that handles query strings and performs semantic-flavored query reformulation for document retrieval, iii) SASKIA, a knowledge base for all knowledge resources and stored data, iv) an indexer that generates standard term index and semantic indexes for all extracted NEs, and v) the LGTE (Lucene with GeoTemporal Extensions), a retrieval and ranking module (Machado, 2009). As knowledge resources, REMBRANDT needs local copies of Wikipedia snapshots (article texts and SQL dumps), an access point (local or remote) to a DBpedia dataset, and access to Yahoo!'s web-service for the geographic ontology GeoPlanet^{TM} (Yahoo!, 2010).

REMBRANDT'S classification strategy begins by mapping NEs to their corresponding Wikipedia and DBpedia pages, using DBpedia's ontology classes and Wikipedia categories to infer the semantic classification (Auer et al., 2007). Then, REMBRANDT applies a set of manually generated language-dependent rules, which represent the internal and external evidence for NEs for a given language, as in "*city of X*" or "*X, Inc.*" This set of rules disambiguates NEs with more than one semantic classification and classifies NEs that were not mapped to a Wikipedia or DBpedia page.

3. User interface

RENOIR is the front-end module for document retrieval, and it can perform semantic-flavored query reformulations. The initial goal of RENOIR was to capture geographic terms in query strings, so that the query's geographic scope could be detected, and be used in the geographic information retrieval step.

In a similar way as REMBRANDT, RENOIR also evolved and expanded its semantic capabilites so that the whole query string could be parsed for semantic clues for the user's intention. As such, RENOIR uses manually-added pattern rules to handle queries, namely query types (what, where, ...), subjects (as in "companies") and conditions (as in "founded in California" and "founded after 1980"). With the captured information, RENOIR estimates the expected answer types (EAT) and selects a reasoning strategy to reformulate the query. For example,



forest fires in Portugal

Results 1 – 10 of 8935 for **forest fires in Portugal.** More info >>

SPAIN SCOLDS CARELESS PUBLIC ABOUT FOREST FIRES

As Spain emerges from its worst summer of forest fires in a decade , the government has begun scolding the public for the careless and sometimes malicious behavior that specialists say causes the vast majority of fires here .

Officials have been uncharacteristically blunt in recent weeks in emphasizing that more than … of the forest fires here are started by people , and that the problem will never be controlled until Spaniards begin to respect …

" I am not going to apologize for saying that … in these fires , " Cristina Narbona , the minister of environment , told reporters late last month . " The ones that should apologize are the people that produce , tolerate and consent to these fires . "

About 12 23,000 forest fires had been recorded in Spain by last week , an increase of more than 25 percent over … last year , according to the Ministry of Environment .

The flames have destroyed more than 12 370,000 acres , an increase of 20 percent , and killed 12 17 people .

Spain endures the most destructive blazes in all of Europe , with nearly half the forest fires that occur each year in the European Union , according to a study by the Confederation of Workers' Commissions , a labor group that represents a large part of the government workers in the

Figure 3: SASKIA user interface

RENOIR can query DBpedia for entities that match the given criteria, and add tem into the initial query string.

SASKIA stores the semantic information from the documents annotated by REMBRANDT, to prepare the index generation. Each NE is stored in a table, together with its terms and semantic classification. NEs successfully grounded to a DBpedia resource are also stored in an ENTITY table, and of those who also were grounded with a geographic place ID, into a GEOSCOPE table as well. SASKIA's data can afterwards be queried for several IE tasks; for example, indexing the geographic NEs of a document collection, to approximate the geographic area of interest of documents, and provide geographic reasoning which retrieving documents for a query with a geographic scope.

Figure presents the RENOIR interface to the LGTE search engine. In the example query "forest fires in Portugal", Portugal is recognised as a geographic entity, and such information is passed into the search engine so that documents with a geographic area of interest within the Portuguese territory have a higher ranking score. This is possible because all documents were previously annotated by REMBRANDT, their NEs stored in SASKIA, and the retrieval engine is capable of reason over the geographic domain on the retrieval step, knowing for example that Porto is a city in Portugal and thus it is within the desired geographic scope.

Figure presents the SASKIA user interface showing an annotated document. Documents can be manually curated by several users, where NEs can be added or removed, change its semantic classification and its grounded information. Documents have a version control system where changes are stored in a different table. Thus, users can work simultaneously in the same document with their own changes, until the collection admin commits those changes permanently to the document. The document version control allows therefore conflict management and undo capabilities to the collection.

Changes can also be made to the stored NEs and grounded information, which impacts all document collections. For instance, if a given entity is not properly classified by REMBRANDT in every document, a single curation in the NE table will therefore trigger a document change on all affected documents. The framework also has a permission control so that users can have read, write or admin permissions to its collections, making it possible to distribute curation work among documents.

4. Conclusion

The REMBRANDT framework is now a mature tool, free and open source software, that can be therefore used by the NLP community on several IE-related tasks. REMBRANDT started as a single NER tool specially crafted to be used for geographic information retrieval systems. As the GIR system evolved while participating into more semantic-flavored evaluation tasks, it also evolved as a full NER framework where rich semantic information extracted from documents is stored, grounded and organised for subsequent processing.

This framework is also adapted to allow curation of its collections and extracted data by several users, in a collaborative and distributive way. With such curated information, the REMBRANDT NER tool can be further improved to take advantage of all previously annotated data and improve its performance specially on the hard, ambiguous NEs.

Acknowledgments

This work is supported by FCT for its LASIGE Multi-annual support, GREASE-II project (grant PTDC/EIA/73614/2006) and a PhD scholarship grant SFRH/BD/45480/2008, and by the Portuguese Government, the European Union (FEDER and FSE) through the Linguateca project, under contract ref. POSC/339/1.3/C/NAC, UMIC and FCCN.

5. References

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. DBpedia: A Nucleus for a Web of Open Data. In Karl Aberer, Key-Sun Choi, Natasha Noy, Dean Allemang, Kyung-Il Lee, Lyndon Nixon, Jennifer Golbeck, Peter Mika, Diana Maynard, Riichiro Mizoguchi, Guus Schreiber, and Philippe Cudré-Mauroux, editors, *6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11–15, 2007, Proceedings*, number 4825 in LNCS, pages 722–735. Springer.

Nuno Cardoso and Mário J. Silva. 2010a. Experiments with Semantic-flavored Query Reformulation of Geo-Temporal Queries. In *Working Notes of the 8th NTCIR Workshop*, Tokyo, Japan, June 15–18.

Nuno Cardoso and Mário J. Silva. 2010b. A GIR Architecture with Semantic-flavored Query Reformulation. In *6th Workshop of Geographic Information Retrieval, GIR 10*, Zurich, Switzerland, 18-19 February.

Nuno Cardoso, David Cruz, Marcirio Chaves, and Mário J. Silva. 2008. Using Geographic Signatures as Query and Document Scopes in Geographic IR. In Carol Peters et al., editors, *Evaluation of Multilingual and Multi-modal Information Retrieval (CLEF 2007). Revised Selected Papers*, LNCS. Springer.

Nuno Cardoso, David Baptista, Francisco J. Lopez-Pellicer, and Mário J. Silva. 2010. Where in the Wikipedia is that answer? the XLDB at the GikiCLEF 2009 task. In Carol Peters, Giorgio Di Nunzio, Mikko Kurimo, Thomas Mandl, Djamel Mostefa, Anselmo Peñas, and Giovanna Roda, editors, *Multilingual Information Access Evaluation Vol. I: Text Retrieval Experiments*. Springer. to appear.

Nuno Cardoso. 2008. REMBRANDT - Reconhecimento de Entidades Mencionadas Baseado em Relações e Análise Detalhada do Texto. In Cristina Mota and Diana Santos, editors, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: Actas do Encontro do Segundo HAREM*, 7 September. In Portuguese.

Jorge Machado. 2009. LGTE: Lucene Extensions for Geo-Temporal Information Retrieval. In *Workshop on Geographic Information on the Internet Workshop (GIIW), held at ECIR 2009*, Toulouse, France, 9 April.

Diana Santos, Paula Carvalho, Hugo Oliveira, and Cláudia Freitas. 2008. Second HAREM: new challenges and old wisdom. In António Teixeira, Vera Lúcia Strube de Lima, Luís Caldas de Oliveira, and Paulo Quaresma, editors, *Computational Processing of Portuguese Language, 8th International Conference (PROPOR'2008), September 8-10, Aveiro, Portugal. Proceedings*, number 5190 in LNCS, pages 212–215. Springer.

Mário J. Silva, Bruno Martins, Marcirio Chaves, Ana Paula Afonso, and Nuno Cardoso. 2006. Adding Geographic Scopes to Web Resources. *CEUS - Computers, Environment and Urban Systems*, 30:378–399.

Yahoo! 2010. Geoplanet^{TM}. `http://developer.yahoo.com/geo/geoplanet/`. Accessed on May 2010.