SeRELeP-Olympics: hot topics for a news portal based on semantic types and named entities

Mírian Bruckschen
Universidade do Vale do Rio dos Sinos
Av Unisinos 905
São Leopoldo, Brazil
mirian.bruckschen@gmail.com

Renata Vieira
Pontifícia Universidade Católica do Rio Grande do Sul
Av Ipiranga 6681
Porto Alegre, Brazil
renata.vieira@pucrs.br

Sandro Rigo
Universidade do Vale do Rio dos Sinos
Av Unisinos 905
São Leopoldo, Brazil
rigo@unisinos.br

ABSTRACT
This paper proposes the use of semantic relations between named entities for identifying hot topics for a news portal on the Olympics subject.

Categories and Subject Descriptors
H.5.4 [**Hypertext/Hypermedia**]: Navigation; H.3.3 [**Information Search and Retrieval**]: Relevance feedback; I.2.7 [**Natural Language Processing**]: Text analysis

1. INTRODUCTION
Information is available virtually everywhere nowadays. When the Internet access became popular, people started to face the other side of information availability: the overload.

In this context, techniques and resources are developed and tested aiming at getting to the important and relevant information. Information extraction systems, recommendation systems, listing of popular or relevant topics or articles, and automatic summarization are just some examples [9].

This paper presents an ongoing project which comprehends a practical application on hot topics' identification in a news portal on the Olympics subject based on a system for automatic identification of semantic relations between entities in texts.

The following sections are distributed in this manner: Section 2 presents related concepts and previous work on the subject; Section 3 introduces the news portal Olympicks.net; Section 4 presents the project, its status, and the semantic processing module it is based on; and, finally, Section 5 ends the paper with final remarks and considerations.

2. RELATED WORK
The Semantic Web and its Web Science proposal is an emerging topic among researchers and the industry [3, 1]. The Web today has the power we know to provide information and to be a way for people to interact, no matter the distance between them. But the Web we know has also some weaknesses, such as the difficulty for computers to get meaning of it. Web pages are structured in a way only humans can read. So, the Semantic Web is not a new Web, but a Web with meaning also for computers.

For this idea to come actually true, it is necessary to build resources for it, such as ontologies, which are structured knowledge bases, and automatic reasoning systems. Systems which automatically extract information from the Web are very important tools now, in this process of making the Semantic Web feasible. Extracting semantic relations between entities in texts, and defining if one entity is the same as other is an important question yet to be solved [2].

Other hot topic in the Web area today is a number of different features, that were given the name of Web2.0. It can be described as an emphasys in the social practice rather than in technical development, and the fact that users are the content producers of dynamic Web pages, instead of the static old ones, built by webmasters. All this social interaction results in many different and new ways of entering and displaying the information: blogs, wikis, mashups, social network websites, social bookmarking, social everything.

Despite its focus on the social aspect, there are some technical features that made Web2.0 possible and really popular. Javascript techniques (AJAX[1]), facilities to data integration from different sources (mashups), possible synergy with the Semantic Web and Natural Language Processing are some examples [7, 6, 8].

Natural Language Processing (NLP) provides the means and resources for automatically processing information available in natural language, english or any other. There are many subareas in this topic, but the area is commonly known by automatic summarization and translation tools, dictionary and other data sets building based on raw text data, text mining and information extraction based on linguistic rules or statistical methods [12].

In the work presented in this paper, its role is the automatic extraction of semantic relations between entities. This information refine the results for a hot topics list in Olympicks.net

[1]Asynchronous JavaScript And XML

news portal, as we will discuss later in this document.

3. OLYMPICKS.NET

Olympicks.net is a news portal on the Olympics subject. The news presented in this portal come from RSS[2] feeds of several portuguese-language sports portals or online newspapers, through Yahoo! Pipes[3]. It has been primarily developed by Digital Communication undergraduate students in a practical course of content production and distribution in the Web. Figure 1 illustrates the portal.

Some of the resources currently available in the news portal are: the retrieval and automatic selection of news of the Olympics, divided into several categories, such as preparatives, modalities and politics; exclusive podcasts on the subject, produced by Olympicks.net team; Flickr[4] photos from the event, geotagged and presented directly in a Yahoo! map; and, finally, trends in the news, which is the part of the portal the SeRELeP-Olympics project refers to.

These trends are, in short, the most frequent topics the news talk about. In Olympics, there are several common topics, but which are the most frequent? Which are the athletes that stand out in the event? The most loved sports? And the most polemical happenings? Olympicks.net trends intends to address these questions by discovering which is the subject for each new item retrieved, and then make a ranking of all the subjects dinamically, presenting them in a tag cloud.

However, identifying hot topics is not an obvious task. There are many ways of doing that, and some of them does not bring the best results. The idea of using NLP resources for making this portal more interesting came from an identified problem with this hot topics feature in Olympicks.net. Since this information was previously retrieved solely based in the tags attributed from the source website, it was not homogeneous or detailed. One of the source websites, for instance, tagged all the news as "Beijing 2008", which does not give any information on the topic the news really talked about. In news that talks about the victory of Michael Phelps, good tags would be the athlete's name itself, "swimming" or "gold medal", for instance. Those give us information on the news subject, and can be used to elect real hot topics. So, we propose to get this information in another manner, which is discussed in this work.

4. SERELEP-OLYMPICS

The system SeRELeP-Olympics intends to use semantic relations between named entities in a portal on the Olympics, the Olympicks.net. The idea behind the system is to use identity, inclusion and occurrence relations between named entities, cited in the text of the news, and stablish hot or common topics. Besides that, as a next step, it is intended to automatically classify the presented news and rank the most relevant to specific searches or the ones referring to favorite topics, by all users, as in a recommendation-like system. Both the working portal and SeRELeP-Olympics proposal are presented in detail in this section.

[2]Really Simple Syndication
[3]`http://pipes.yahoo.com/pipes/`
[4]`http://flickr.com`

4.1 SeRELeP

SeRELeP is a tool for the recognition of relations between named entities in texts. Using linguistic rules, it aims to recognize relations of identity, inclusion and occurrence between previously identified entities. SeRELeP and its auxiliary package SeRELeP Tools were influenced by HAREM[5] directives [10, 11]. The relations it intends to extract are some of the ones proposed by HAREM, and so are the rules for them to be extracted. It uses the parser PALAVRAS [4] for the identification of named entities, and the Tiger2XCES [5] conversor from PALAVRAS TigerXML format to the XML CES format[6]. Its pipeline is illustrated in Figure 2.

The relation of identity is attributed to entities which refer to the same object in the world (such as United Nations, UN and a different occurrence of UN in the same text). Obviously, it depends on the classification of the entity; if Brazil referres to the soccer team, it is not a place, and therefore should not have the identity relation attributed to another occurrence of Brazil in the text referring to Brazil as the country of a very beautiful coast.

The other relations treated by SeRELeP are inclusion and occurrence. The first is attributed to geographical entities, as long as one includes the other (like the USA includes Idaho, which includes Boise). And, finally, the occurrence relation is attributed to entities of events or organizations and places (as in an organization which is based somewhere, or an event which occurs in a place; an example would be Olympics occurring in China).

This system participated in HAREM 2008 and got promising results[7], showing that it would be already possible to try it in a real world, practical application. It is important to remember, however, that these results refer to the processing and analysis of HAREM corpus, which was very heterogeneous (jornalistic texts, pages directly retrieved from the Web, interviews, and so on). A different situation occurs with SeRELeP-Olympics, since it is supposed to process only jornalistic texts. Yet, it is intended to improve this results using specific-domains ontologies and information from the Web (initially, from Wikipedia website).

4.2 Olympicks.net and SeRELeP Integration

SeRELeP-Olympics relies on the retrieval of news from the very same RSS feed, generated by Yahoo! Pipes, the current version of Olympicks.net portal uses to display its content. The pipeline of the whole SeRELeP-Olympics project is illustrated in Figure 3.

After getting these news and processing them, SeRELeP-Olympics should return a list of hot topics. This feed with the list of hot topics should be used in Olympicks.net news portal instead of the current algorithm for calculating the frequency in which the news tags appear.

The main idea is based in the processing of the retrieved news and extraction of semantic relations and other syntac-

[5]HAREM is a joint evaluation for named entities recognizers
[6]XML Corpus Encoding Standard, as available at `http://www.xces.org/`
[7]`http://linguateca.dei.uc.pt/harem/resultados`

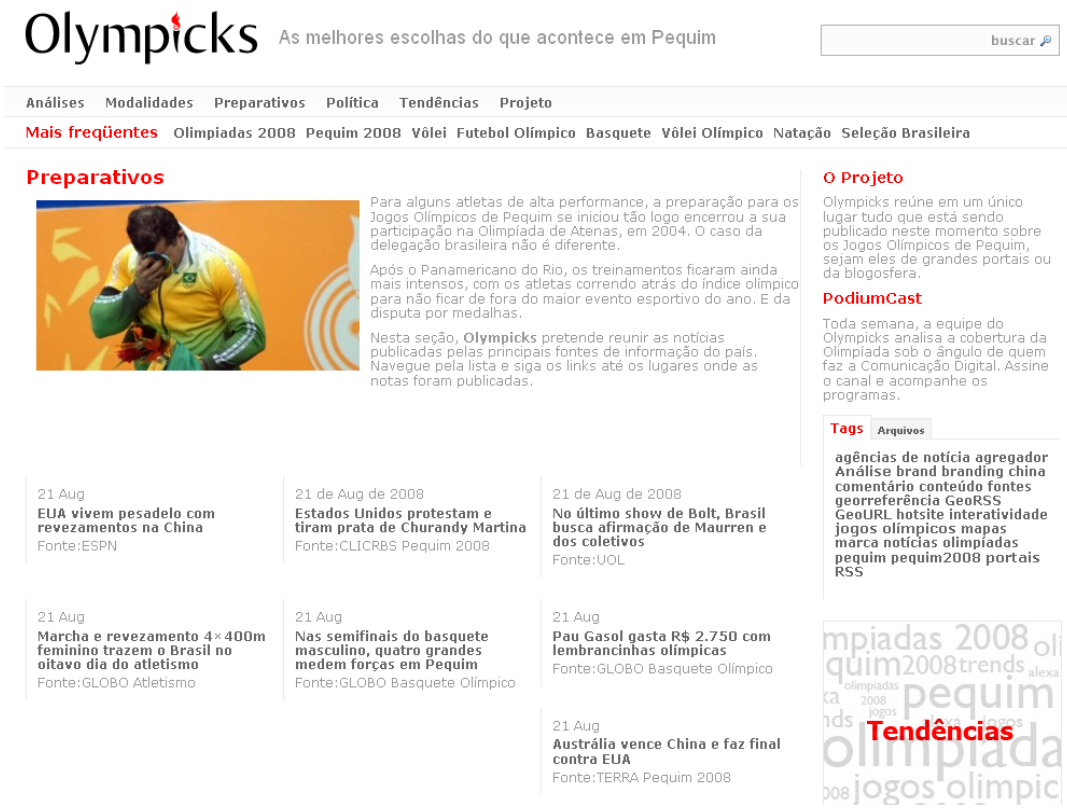

Figure 1: Olympicks.net news portal

tic and semantic information. The parser PALAVRAS provides us information in syntactic and semantic levels, both useful in this work. It classifies the various elements of the phrases and returns to us where is the main verb, the subject and predicate, if there is any apposition, and so on. Also, it marks many processed tokens, mostly the nouns, with one tag from a prebuilt categories set: human (individual, groups), professions, biology-related, sports-related, organizations, and a lot others. Using syntactic information, we can select some specific syntactic classes for observation: (proper) nouns and adjectives, mostly. In the other hand, using semantic information provided by PALAVRAS, we can retrieve only relevant items from those (nouns referring to sports and human professions and nationality adjectives for instance).

After that, checking the frequency of these items in the news gives us a more accurate hot topics' list than the previously attributed tags. This approach would be very useful, also, for automatic clustering and news tagging.

At first, we propose using identity relation between entities to improve the hot topics list. All news talking about "Cielo", "Cesar Cielo" and "Cesar Cielo Filho" relate to the same entity, which represents the brazilian freestyle swimmer who won the gold medal for Brazil. All of them, so, should increase the frequency of this entity, giving Cielo a chance to appear in Olympicks.net hot topics. If we let these occurrences alone, he probably never would, since there are many ways of referring to the same entity. Our proposed identity relation treatment intends to solve this issue in SeRELeP-Olympics.

Besides the relations, other items that should be object of evaluation are the categories selected for candidates to hot topic. As it is based in a totally automatic processing, it is likely that some identified terms are not really good candidates for a hot topic (as "changes", a word that often appears in political news on the "Olympics on China" subject, for instance). So, it is necessary also the creation and maintenance of stopwords beside the usual (prepositions, adverbs, articles).

5. FINAL REMARKS

This paper presented a proposal for integration of NLP techniques in Olympicks.net, a news portal on the Olympics subject. By automatically extracting semantic relations between the entities which appear in the news texts, it is intended to elect better hot topics for the Web portal.

As future work, we plan to use other relations besides identity. The inclusion of Beijing in China, for instance, indicates that news which occur in Beijing occur also in China, and therefore should be tagged so. The same for occurrence relations, and other that can be included in SeRELeP in the future.

Other applications are intended to be developed using the same approach, like automatically ranking most relevant news, automatic categorization of news and articles, and

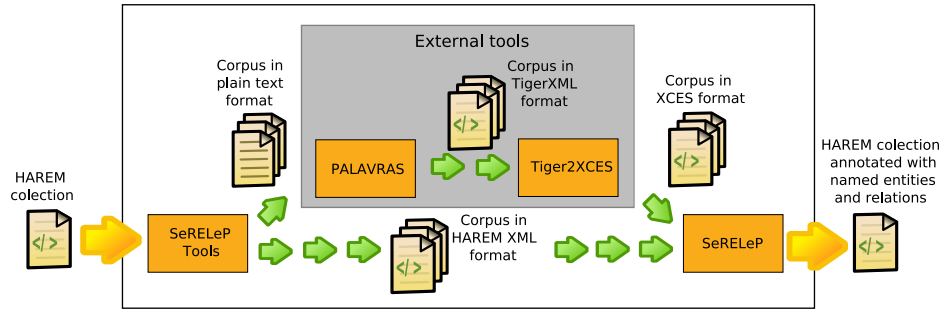

Figure 2: Pipeline for the automatic annotation of relations in HAREM colection

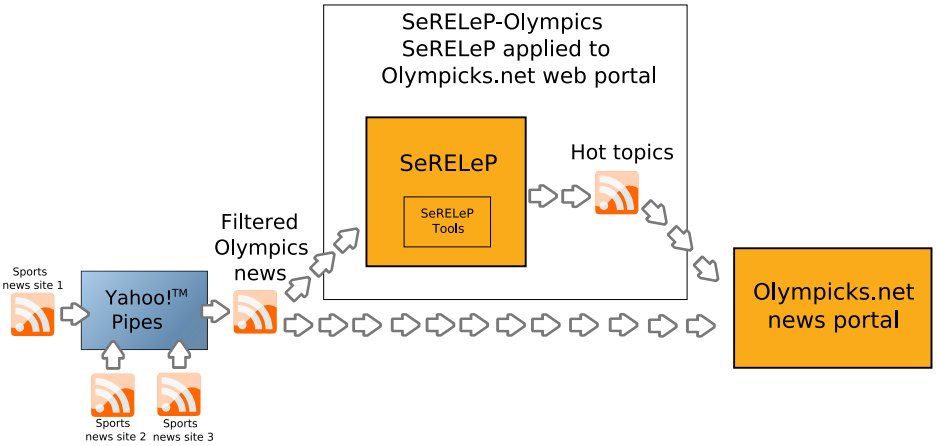

Figure 3: Proposed pipeline for SeRELeP-Olympics project

identification of relations and similarity between different texts.

6. ACKNOWLEDGEMENTS

We would like to thank CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) and FAPERGS (Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul) for their support in this project.

7. REFERENCES

[1] H. Alani, P. Chandler, W. Hall, K. O'Hara, N. Shadbolt, and M. Szomszor. Building a pragmatic Semantic Web. *IEEE Intelligent Systems*, 23(3):61–68, 2008.

[2] T. Berners-Lee, W. Hall, J. A. Hendler, K. OHara, N. Shadbolt, and D. J. Weitzner. A framework for Web Science. *Foundations and Trends in Web Science*, 1(1):1–130, 2006.

[3] T. Berners-Lee, J. Hendler, and O. Lassila. The Semantic Web (berners-lee et. al 2001). *Scientific American*, May 2001.

[4] E. Bick. *The Parsing System PALAVRAS – Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. PhD thesis, Department of Linguistics, University of Århus, Dinamarca, 2000.

[5] J. G. C. de Souza. Tiger2XCES: Software de conversão de arquivos no formato TigerXML para formato XCES, 2007.

[6] D. E. Millard and M. Ross. Web 2.0: hypertext by any other name? In *HYPERTEXT '06: Proceedings of the seventeenth conference on Hypertext and hypermedia*, pages 27–30, New York, NY, USA, 2006. ACM Press.

[7] L. J. B. Nixon. Multimedia, web 2.0 and the semantic web: A strategy for synergy. In *Semantic Web for Multimedia Annotation workshop, WWW 2006 conference*, 2006.

[8] T. I. M. Oreilly. What is web 2.0: Design patterns and business models for the next generation of software. *Social Science Research Network Working Paper Series*, 2007.

[9] T. Pardo. Gistsumm - gist summarizer: Extensões e novas funcionalidades. Technical report, NILC-TR-05-05. São Carlos-SP, 2005.

[10] D. Santos and N. Cardoso. A golden resource for named entity recognition in portuguese. In *7th Workshop on Computational Processing of Written and Spoken Language (PROPOR'2006)*, Itatiaia, RJ, Brasil, 2006. LNAI 3960, Springer.

[11] D. Santos and N. Cardoso, editors. *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*. Linguateca, Portugal, 2007.

[12] R. Vieira and V. L. S. de Lima. JAIA/Linguística Computacional: Princípios e aplicações. In A. T. Martins and D. L. Borges, editors, *As Tecnologias da informação e a questão social: anais*, Fortaleza, CE, Brasil, 2001.