A Comparable Corpus Based on Aligned Multilingual Ontologies

Roger Granada
PUCRS (Brazil)
roger.granada@acad.pucrs.br

Lucelene Lopes
PUCRS (Brazil)
lucelene.lopes@pucrs.br

Carlos Ramisch
University of Grenoble (France)
ceramisch@inf.ufrgs.br

Cassia Trojahn
University of Grenoble (France)
cassia.trojahn@inria.fr

Renata Vieira
PUCRS (Brazil)
renata.vieira@pucrs.br

Aline Villavicencio
UFRGS (Brazil)
alinev@gmail.com

Abstract

In this paper we present a methodology for building comparable corpus, using multilingual ontologies of a scpecific domain. This resource can be exploited to foster research on multilingual corpus-based ontology learning, population and matching. The building resource process is exemplified by the construction of annotated comparable corpora in English, Portuguese, and French. The corpora, from the conference organization domain, are built using the multilingual ontology concept labels as seeds for crawling relevant documents from the web through a search engine. Using ontologies allows a better coverage of the domain. The main goal of this paper is to describe the design methodology followed by the creation of the corpora. We present a preliminary evaluation and discuss their characteristics and potential applications.

1 Introduction

Ontological resources provide a symbolic model of the concepts of a scientific, technical or general domain (e.g. Chemistry, automotive industry, academic conferences), and of how these concepts are related to one another. However, ontology creation is labour intensive and error prone, and its maintenance is crucial for ensuring the accuracy and utility of a given resource. In multilingual contexts, it is hard to keep the coherence among ontologies described in different languages and to align them accurately. These difficulties motivate the use of semi-automatic approaches for cross-lingual ontology enrichment and population, along with intensive reuse and interoperability between ontologies. For that, it is crucial to have domain-specific corpora available, or the means of automatically gathering them.

Therefore, this paper describes an ontology-based approach for the generation of multilingual comparable corpora. We use a set of multilingual domain-dependent ontologies, which cover different aspects of the conference domain. These ontologies provide the seeds for building the domain specific corpora from the web. Using high-level background knowledge expressed in concepts and relations, which are represented as natural language descriptions in the labels of the ontologies, allow focused web crawling with a semantic and contextual coverage of the domain. This approach makes web crawling more precise, which is crucial when exploiting the web as a huge corpus.

Our motivation is the need of such resources in tasks related to semi-automatic ontology creation and maintenance in multilingual domains. We exemplify our methodology focusing on the construction of three corpora, one in English, one in Portuguese, and one in French. This effort is done in the context of a larger research project which aims at investigating methods for the construction of lexical resources, integrating multilingual lexica and ontologies, focusing on collaborative and automatic techniques (http://cameleon.imag.fr/xwiki/bin/view/Main/).

In the next section, we present some relevant related work (§2). This is followed by a description of the methodology used to build the corpora (§3). Finally, the application example expressed by the resulting corpora are evaluated (§4) and discussed

Proceedings of the First Workshop on Multilingual Modeling, pages 25–31,
Jeju, Republic of Korea, 8-14 July 2012.

(§5). We conclude by outlining their future applications (§ 6).

2 Related Work

Web as corpus (WAC) approaches have been successfully adopted in many cases where data sparseness plays a major limiting role, either in specific linguistic constructions and words in a language (e.g. compounds and multiword expressions), or for less resourced languages in general[1].

For instance, Grefenstette (1999) uses WAC for machine translation of compounds from French into English, Keller et al. (2002) for adjective-noun, noun-noun and verb-object bigram discovery, and Kim and Nakov (2011) for compound interpretation. Although a corpus derived from the web may contain noise, the sheer size of data available should compensate for that. Baroni and Ueyama (2006) discuss in details the process of corpus construction from web pages for both generic and domain-specific corpora. In particular, they focus on the cleaning process applied to filter the crawled web pages. Much of the methodology applied in our work is similar to their proposed approach (see §3).

Moreover, when access to parallel corpora is limited, comparable corpora can minimize data sparseness, as discussed by Skadina et al. (2010). They create bilingual comparable corpora for a variety of languages, including under-resourced ones, with 1 million words per language. This is used as basis for the definition of metrics for comparability of texts. Forsyth and Sharoff (2011) compile comparable corpora for terminological lexicon construction. An initial verification of monolingual comparability is done by partitioning the crawled collection into groups. Those are further extended through the identification of representative archetypal texts to be used as seeds for finding documents of the same type.

Comparable corpora is a very active research subject, being in the core of several European projects (e.g. TTC[2], Accurat[3]). Nonetheless, to date most of the research on comparable corpora seems to focus on lexicographic tasks (Forsyth and Sharoff, 2011; Sharoff, 2006), bilingual lexicon extraction (Morin and Prochasson, 2011), and more generally on machine translation and related applications (Ion et al., 2011). Likewise, there is much to be gained from the potential mutual benefits of comparable corpora and ontology-related tasks.

Regarding multilingually aligned ontologies, very few data sets have been made available for use in the research community. Examples include the vlcr[4] and the mldirectory[5] datasets. The former contains a reduced set of alignments between the thesaurus of the Netherlands Institute for Sound and Vision and two other resources, English WordNet and DBpedia. The latter consists of a set of alignments between web site directories in English and in Japanese. However, these data sets provide subsets of bilingual alignments and are not fully publicly available. The MultiFarm dataset[6], a multilingual version of the OntoFarm dataset (Šváb et al., 2005), has been designed in order to overcome the lack of multilingual aligned ontologies. MultiFarm is composed of a set of seven ontologies that cover the different aspects of the domain of organizing scientific conferences. We have used this dataset as the basis for generating our corpora.

3 Methodology

The main contribution of this paper is the proposal of the methodology to build corpora. This section describes the proposed methodology presenting our own corpus crawler, but also its application to construct three corpora, in English, Portuguese, and French. These corpora are constructed from the MultiFarm dataset.

3.1 Tools and Resources

Instead of using an off-the-shelf web corpus tool such as BootCaT (Baroni and Bernardini, 2004), we implemented our own corpus crawler. This allowed us to have more control on query and corpus construction process. Even though our corpus construc-

[1] Kilgarriff (2007) warns about the dangers of statistics heavily based on a search engine. However, since we use the downloaded texts of web pages instead of search engine count estimators, this does not affect the results obtained in this work.

[2] www.ttc-project.eu

[3] www.accurat-project.eu

[4] www.cs.vu.nl/~laurah/oaei/2009

[5] oaei.ontologymatching.org/2008/mldirectory

[6] web.informatik.uni-mannheim.de/multifarm

tion strategy is similar to the one implemented in BootCaT, there are some significant practical issues to take into account, such as:

- The predominance of multiword keywords;
- The use of the fixed keyword *conference*;
- The expert tuning of the cleaning process;
- The use of a long term support search AP[b].

Besides, BootCaT uses the Bing search API, which will no longer work in 2012. As our work is part of a long-term project, we preferred to use Google's search API as part of the University Research Program.

The set of seed domain concepts comes from the MultiFarm dataset. Seven ontologies from the OntoFarm project (Table 1), together with the alignments between them, have been translated from English into eight languages (Chinese, Czech, Dutch, French, German, Portuguese, Russian, and Spanish). As shown in Table 1, the ontologies differ in numbers of classes, properties, and in their logical expressivity. Overall, the ontologies have a high variance with respect to structure and size and they were based upon three types of resources:

- actual conferences and their web pages (type 'web'),
- actual software tools for conference organisation support (type 'tool'), and
- experience of people with personal participation in organisation of actual conferences (type 'insider').

Currently, our comparable corpus generation approach focuses on a subset of languages, namely English (`en`), Portuguese (`pt`) and French (`fr`). The labels of the ontology concepts, like *conference* and *call for papers*, are used to generate queries and retrieve the pages in our corpus. In the current implementation, the structure and relational properties of the ontologies were ignored. Concept labels were our choice of seed keywords since we intended to have comparable, heterogeneous and multilingual domain resources. This means that we need a corpus *and* an ontology referring to the same set of terms or concepts. We want to ensure that the concept labels are present in the corresponding natural language, textual sources. This combination of resources is essential for our goals, which involve problems such as ontology learning and enriching from corpus. Thus, the original ontology can serve as a reference for automatically extracted resources. Moreover, we intend to use the corpus as an additional resource for ontology (multilingual) matching, and again the presence of the labels in the corpus is of great relevance.

Name	Type	C	DP	OP
Ekaw	insider	74	0	33
Sofsem	insider	60	18	46
Sigkdd	web	49	11	17
Iasted	web	140	3	38
ConfTool	tool	38	23	13
Cmt	tool	36	10	49
Edas	tool	104	20	30

Table 1: Ontologies from the OntoFarm dataset in terms of number of classes (C), datatype properties (DP) and object properties (OP).

3.2 Crawling and Preprocessing

In each language, a concept label that occurs in two or more ontologies provides a seed keyword for query construction. This results in 49 `en` keywords, 54 `pt` keywords and 43 `fr` keywords. Because many of our keywords are formed by more than one word (average length of keywords is respectively 1.42, 1.81 and 1.91 words), we combine three keywords regardless of their sizes to form a query. The first keyword is static, and corresponds to the word *conference* in each language. The query set is thus formed by permuting keywords two by two and concatenating the static keyword to them (e.g. *conference reviewer program committee*). This results in $1 \times 48 \times 47 = 2,256$ `en` queries, 2,756 `pt` queries and 1,892 `fr` queries. Average query length is 3.83 words for `en`, 4.62 words for `pt` and 4.91 words for `fr`. This methodology is in line with the work of Sharoff (2006), who suggests to build queries by combining 4 keywords and downloading the top 10 URLs returned for each query.

The top 10 results returned by Google's search

API[7] are downloaded and cleaned. Duplicate URLs are automatically removed. We did not filter out URLs coming from social networks or Wikipedia pages because they are not frequent in the corpus. Results in formats other than html pages (like .doc and .pdf documents) are ignored. The first cleaning step is the extraction of raw text from the html pages. In some cases, the page must be discarded for containing malformed html which our page cleaner is not able to parse. In the future, we intend to improve the robustness of the HTML parser.

3.3 Filtering and Linguistic Annotation

After being downloaded and converted to raw text, each page undergoes a two-step processing. In the first step, markup characters as interpunctuation, quotation marks, etc. are removed leaving only letters, numbers and punctuation. Further heuristics are applied to remove very short sentences (less than 3 words), email addresses, URLs and dates, since the main purpose of the corpus is related to concept, instance and relations extraction. Finally, heuristics to filter out page menus and footnotes are included, leaving only the text of the body of the page. The raw version of the text still contains those expressions in case they are needed for other purposes.

In the second step, the text undergoes linguistic annotation, where sentences are automatically lemmatized, POS tagged and parsed. Three well-known parsers were employed: Stanford parser (Klein and Manning, 2003) for texts in English, PALAVRAS (Bick, 2000) for texts in Portuguese, and Berkeley parser (Petrov et al., 2006) for texts in French.

4 Evaluation

The characteristics of the resulting corpora are summarized in tables 2 and 3. Column D of table 2 shows that the number of documents retrieved is much higher in `en` than in `pt` and `fr`, and this is not proportional to the number of queries (Q). Indeed, if we look in table 3 at the average ratio of documents retrieved per query (D/Q), the `en` queries return much more documents than queries in other languages. This indicates that the search engine returns more distinct results in `en` and more duplicate URLs in `fr` and in `pt`. The high discrepancy in the number of documents has a direct impact in the size of the corpus in each language. However, this is counterbalanced by the average longer documents (S/D) and longer sentences (W/S) in `pt` and `fr` with respect to `en`. The raw corpus contains from 9.48 million words in `fr`, 12.88 million words in `pt` to 15.85 million words in `en`, constituting a large resource for research on ontology-related tasks.

	Q	D	W token	W type
en	2,256	10,127	15,852,650	459,501
pt	2,756	5,342	12,876,344	405,623
fr	1,892	5,154	9,482,156	362,548

Table 2: Raw corpus dimensions: number of queries (Q), documents (D), and words (W).

	D/Q	S/D	W/S	TTR
en	4.49	110.59	14.15	2.90%
pt	1.94	120.08	20.07	3.15%
fr	2.72	115.63	15.91	3.82%

Table 3: Raw corpus statistics: average documents per query (D/Q), sentences per document (S/D), words per sentence (W/S) and type-token ration (TTR).

A preliminary semi-automated analysis of the corpus quality was made by extracting the top-100 most frequent n-grams and unigrams for each language. Using the parsed corpora, the extraction of the top-100 most frequent n-grams for each language focused on the most frequent noun phrases composed by at least two words. The lists with the top-100 most frequent unigrams was generated by extracting the most frequent nouns contained in the parsed corpus for each language. Four annotators manually judged the semantic adherence of these lists to the conference domain.

We are aware that semantic adherence is a vague notion, and not a straightforward binary classification problem. However, such a vague notion was considered useful at this point of the research, which is ongoing work, to give us an initial indication of the quality of the resulting corpus. Examples of what we consider adherent terms are *appel á communication (call for papers)*, *conference program* and *texto completo (complete text)*, examples

[7] `research.google.com/university/search`

	# of adherent terms	
	Lower	**Upper**
`en` words	46	85
`en` n-grams	57	94
`fr` words	21	69
`fr` n-grams	24	45
`pt` words	32	70
`pt` n-grams	11	45

Table 4: Number of words and n-grams judged as semantically adherent to the domain.

of nonadherent terms extracted from the corpus were *produits chimiques (chemical products)*, *following case*, *projeto de lei (law project)*. In the three languages, the annotation of terms included misparsed and mistagged words (*ad hoc*), places and dates typical of the genre (but not necessarily of the domain), general-purpose terms frequent in conference websites (*email*, *website*) and person names.

Table 4 shows the results of the annotation. The lower bound considers an n-gram as semantically adherent if all the judges agree on it. The upper bound, on the other hand, considers as relevant n-grams all those for which at least one of the four judges rated it as relevant. As a result of our analysis, we found indications that the English corpus was more adherent, followed by French and Portuguese. This can be explained by the fact that the amount of internet content is larger for English, and that the number of international conferences is higher than national conferences adopting Portuguese and French as their official languages. We considered the adherence of Portuguese and French corpora rather low. There are indications that material related to political meetings, law and public institutions was also retrieved on the basis of the seed terms.

The next step in our evaluation is verifying its comparable nature, by counting the proportion of translatable words. Thus, we will use existing bilingual dictionaries and measure the rank correlation of equivalent words in each language pair.

5 Discussion

The first version of the corpus containing the totality of the raw pages, the tools used to process them, and a sample of 1,000 annotated texts for each language are freely available for download at the CAMELEON project website[8]. For the raw files, each page is represented by an URL, a language code, a title, a snippet and the text of the page segmented into paragraphs, as in the original HTML file. A companion log file contains information about the download dates and queries used to retrieve each URL. The processed files contain the filtered and parsed texts. The annotation format varies for each language according to the parser used. The final version of this resource will be available with the totality of pages parsed.

Since the texts were extracted from web pages, there is room for improvement concerning some important issues in effective corpus cleaning. Some of these issues were dealt with as described in the § 3, but other issues are still open and are good candidates for future refinements. Examples already foreseen are the removal of foreign words, special characters, and usual web page expressions like "site under construction", "follow us on twitter", and "click here to download". However, the relevance of some of these issues depends on the target application. For some domains, foreign expressions may be genuine part of the vocabulary (e.g. *parking* or *weekend* in colloquial French and *deadline* in Portuguese), and as such, should be kept, while for other domains these expressions may need to be removed, since they do not really belong to the domain. Therefore, the decision of whether to implement these filters or not, and to deal with truly multilingual texts, depends on the target application.

Another aspect that was not taken into account in this preliminary version is related to the use of the relations between concepts in the ontologies to guide the construction of the queries. Exploiting the contextual and semantic information expressed in these relations may have an impact in the set of retrieved documents and will be exploited in future versions of the corpus.

6 Conclusions and Future Work

This paper has described an ontology-based approach for the generation of a multilingual compara-

[8] `cameleon.imag.fr/xwiki/bin/view/Main/Resources`

ble corpus in English, Portuguese and French. The corpus constructed and discussed here is an important resource for ontology learning research, freely available to the research community. The work on term extraction that we are doing for the initial assessment of the corpus is indeed the initial step towards more ambitious research goals such as multilingual ontology learning and matching in the context of our long-term research project.

The initial ontologies (originally built by hand) and resulting corpora can serve as a reference, a research resource, for information extraction tasks related to ontology learning (term extraction, concept formation, instantiation, etc). The resource also allows the investigation of ontology enriching techniques, due to dynamic and open-ended nature of language, by which relevant terms found in the corpus may not be part of the original ontology. We can also assess the frequencies (relevance) of the labels of the ontology element with respect to the corpus, thus assessing the quality of the ontology itself. Another research that can be developed on the basis of our resource is to evaluate the usefulness of a corpus in the improvement of existing multilingual ontology matching techniques[9].

Regarding to our own crawler implementation, we plan to work on its evaluation by using other web crawlers, as BootCaT, and compare both approaches, specially on what concerns the use of ontologies.

From the point of view of NLP, several techniques can be compared showing the impact of adopting different tools in terms of depth of analysis, from POS tagging to parsing. This is also an important resource for comparable corpora research, which can be exploited for other tasks such as natural language translation and ontology-based translation. So far this corpus contains English, Portuguese and French versions, but the ontology data set includes 8 languages, to which this corpus may be extended in the future.

[9] An advantage of this resource is that the Multilingual OntoFarm is to be included in the OAEI (Ontology Alignment Evaluation Initiative) evaluation campaign.

References

Marco Baroni and Silvia Bernardini. 2004. BootcaT: Bootstrapping corpora and terms from the web. In *Proc. of the Fourth LREC (LREC 2004)*, Lisbon, Portugal, May. ELRA.

Marco Baroni and Motoko Ueyama. 2006. Building general- and special-purpose corpora by web crawling. In *Proceedings of the 13th NIJL International Symposium on Language Corpora: Their Compilation and Application*, pages 31–40.

Eckhard Bick. 2000. *The parsing system Palavras*. Aarhus University Press.

Richard Forsyth and Serge Sharoff. 2011. From crawled collections to comparable corpora: an approach based on automatic archetype identification. In *Proc. of Corpus Linguistics Conference*, Birmingham, UK.

Gregory Grefenstette. 1999. The World Wide Web as a resource for example-based machine translation tasks. In *Proc. of the Twenty-First Translating and the Computer*, London, UK, Nov. ASLIB.

Radu Ion, Alexandru Ceauşu, and Elena Irimia. 2011. An expectation maximization algorithm for textual unit alignment. In Zweigenbaum et al. (Zweigenbaum et al., 2011), pages 128–135.

Frank Keller, Maria Lapata, and Olga Ourioupina. 2002. Using the Web to overcome data sparseness. In Jan Hajič and Yuji Matsumoto, editors, *Proc. of the 2002 EMNLP (EMNLP 2002)*, pages 230–237, Philadelphia, PA, USA, Jul. ACL.

Adam Kilgarriff. 2007. Googleology is bad science. *Comp. Ling.*, 33(1):147–151.

Su Nam Kim and Preslav Nakov. 2011. Large-scale noun compound interpretation using bootstrapping and the web as a corpus. In *Proc. of the 2011 EMNLP (EMNLP 2011)*, pages 648–658, Edinburgh, Scotland, UK, Jul. ACL.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proc. of the 41st ACL (ACL 2003)*, pages 423–430, Sapporo, Japan, Jul. ACL.

Emmanuel Morin and Emmanuel Prochasson. 2011. Bilingual lexicon extraction from comparable corpora enhanced with parallel corpora. In Zweigenbaum et al. (Zweigenbaum et al., 2011), pages 27–34.

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *Proc. of the 21st COLING and 44th ACL (COLING/ACL 2006)*, pages 433–440, Sidney, Australia, Jul. ACL.

Serge Sharoff, 2006. *Creating general-purpose corpora using automated search engine queries*. Gedit, Bologna, Italy.

Inguna Skadina, Ahmed Aker, Voula Giouli, Dan Tufiş, Robert Gaizauskas, Madara Mieirina, and Nikos Mastropavlos. 2010. A Collection of Comparable Corpora for Under-resourced Languages. In Inguna Skadina and Andrejs Vasiljevs, editors, *Frontiers in Artificial Intelligence and Applications*, volume 219, pages 161–168, Riga, Latvia, Oct. IOS Press.

Ondřej Šváb, Vojtěch Svátek, Petr Berka, Dušan Rak, and Petr Tomášek. 2005. Ontofarm: Towards an experimental collection of parallel ontologies. In *Poster Track of ISWC 2005*.

Pierre Zweigenbaum, Reinhard Rapp, and Serge Sharoff, editors. 2011. *Proc.of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web (BUCC 2011)*, Portland, OR, USA, Jun. ACL.