The CAMELEON comparable corpus for supporting ontology-related tasks

Roger Granada, Lucelene Lopes, Carlos Ramisch, Cassia Trojahn, Renata Vieira and Aline Villavicencio

CAMELEON Workshop 2011

The problem

- Ontologies model concepts of a scientific, technical or general domain (e.g. Chemistry)
- Ontology creation is onerous
- Domains and language are dynamic

What to do?

Semi-automatic web-based methods :-)

Goals

- Ontology-based approach for the generation of a multilingual comparable corpus
- Ontology $\rightarrow$ labels $\rightarrow$ keywords $\rightarrow$ web documents $\rightarrow$ text

Motivation

- Useful research resource to develop methods for supporting tasks related to semi-automatic ontology creation and maintenance
- Target languages: English, Portuguese and French.

Web as corpus

- Machine translation of compounds

  [Grefenstette, 1999]
- Compound interpretation [Kim and Nakov, 2011]
- MWE extraction [Keller et al., 2002]

Web corpus construction

- Methodology for keyword selection, cleaning, etc. [Baroni and Ueyama, 2006]
- Googleology is bad science [Kilgarriff, 2007]

Comparable corpora

- Minimize data sparseness in absence of parallel data
- Bilingual lexicon construction

  [Skadina et al., 2010, Morin and Prochasson, 2011]
- Terminology construction [Forsyth and Sharoff, 2011], lexicography [Sharoff, 2006]
- `www.ttc-project.eu` and `www.accurat-project.eu`

Multilingual ontologies

- vlcr dataset `www.cs.vu.nl/~laurah/oaei/2009`
- mldirectory dataset `oaei.ontologymatching.org/2008/mldirectory`
- MultiFarm dataset `web.informatik.uni-mannheim.de/multifarm`
- In general, lack of multilingual aligned ontologies

Methodology

- Similar to BootCaT [Baroni and Bernardini, 2004] - control process
- Seeds: The labels of OntoFarm ontology concepts. e.g. *conference* and *call for papers*
- Concept labels that occur in 2+ ontologies $\rightarrow$ seed keyword for query construction
- 49 `en` keywords, 54 `pt` keywords and 43 `fr` keywords

Ontologies

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

Crawling

- 3 keywords $\rightarrow$ query
- 1st keyword static: *conference*
- *e.g. conference reviewer program committee*
- *Top-10 non duplicate URLs with Google*
- *Only proper HTML, no URL filtering (wikipedia, FB)*

Cleaning

- Markup characters: interpunctuation, quotation marks, etc.
- Short sentences (less than 3 words)
- Email addresses, URLs and dates
- Page menus and footnotes

Annotation

- Stanford parser [Klein and Manning, 2003] for English
- PALAVRAS [Bick, 2000] for Portuguese
- Berkeley parser [Petrov et al., 2006] for French

Corpus characteristics

	Q	D	S	W token	W type
en	2,256	10,127	1,119,971	15,852,650	459,501
pt	2,756	5,342	641,452	12,876,344	405,623
fr	1,892	5,154	595,938	9,482,156	362,548

Corpus characteristics

	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%

Future work

- Improved cleaning
- Truly multilingual documents
- Information extraction, bilingual lexica
- Ontology enrichment, population, alignment
- `cameleon.imag.fr/xwiki/bin/view/Main/Resources`

Thank you! Questions?

Roger Granada, Lucelene Lopes, Carlos Ramisch, Cassia Trojahn, Renata Vieira and Aline Villavicencio

CAMELEON Workshop 2011

Thank you! Questions?

References II

Baroni, M. and Bernardini, S. (2004).

BootcaT: Bootstrapping corpora and terms from the web.

In *Proc. of the Fourth LREC (LREC 2004)*, Lisbon, Portugal. ELRA.

Baroni, M. and Ueyama, M. (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.

Bick, E. (2000).

The parsing system Palavras.

Aarhus University Press.

Forsyth, R. and Sharoff, S. (2011).

From crawled collections to comparable corpora: an approach based on automatic archetype identification.

In *Proc. of Corpus Linguistics Conference*, Birmingham, UK.

Grefenstette, G. (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. ASLIB.