Preliminary results of the Cameleon shared task on ontology translation

Roger Granada, Carlos Ramisch, Cassia Trojahn

4th Cameleon Workshop

December 13, 2013

Outline

Outline

Objectives

1. Focus on a task that is of common interest within Cameleon

2. Experiment different strategies for ontology translation
   - No restriction on the kind of method, approach or tool
   - No restriction on the use of external resources

Task description

- Provide automatic translations for a set of ontologies in EN, FR and PT :
    - EN ontology $\rightarrow$ FR ontology
    - EN ontology $\rightarrow$ PT ontology
    - FR ontology $\rightarrow$ EN ontology
    - FR ontology $\rightarrow$ PT ontology
    - PT ontology $\rightarrow$ EN ontology
    - PT ontology $\rightarrow$ FR ontology

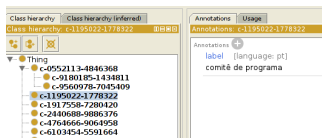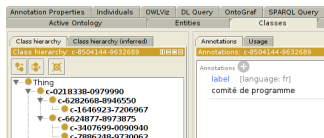- Provide manual translations

Outline

MultiFarm data set

- 7 manually created ontologies of the conference domain
- Ontologies manually translated into 8 languages:
  - CZ, CN, DE, ES, FR, NL, PT, RU (+EN)
- From manually created alignments between original ontologies, cross-lingual alignments were automatically generated

Task data set

- Shared task: 5 ontologies in EN, FR and PT
  - 3 ontologies for open tests (cmt,conference,confOf)
  - 2 ontologies for blind tests (edas,ekaw)

- Ontologies available in OWL (Web Ontology Language) format

Manual translation examples

hasCo-author	possui pedido de revisão
hasConferenceMember	possui participante da conferência
hasConflictOfInterest	possui conflito de interesses
hasDecision	possui decisão
hasProgramCommitteeMember	possui um membro do comitê de programa
hasSubjectArea	possui uma área
markConflictOfInterest	marca um conflito de interesse
memberOfConference	membro da conferência
memberOfProgramCommittee	membro do comite de programa
paperAssignmentFinalizedBy	atribuição de um artigo finalizada por
paperAssignmentToolsRunBy	ferramentas de atribuião de artigos executadas por
printHardcopyMailingManifests	impressão de documentos de envio de volume de correio
readByMeta-Reviewer	leitura realizada por um meta revisor

Statistics of ontologies

Ontology	Lang	#Labels	1-gram	2-gram	3-gram	>3-gram
Cmt	en	88	24	26	18	20
	fr	88	23	8	20	37
	pt	88	15	23	16	34
Conference	en	123	18	38	30	37
	fr	123	16	14	34	59
	pt	123	18	27	38	40
ConfOf	en	74	24	17	16	17
	pt	74	25	5	18	26
	fr	74	25	19	10	20
Total		855	188	177	200	290

Outline

Evaluation

- Based on the comparison of provided translations with manually generated translations :
  1. exact match (exact translation)
  2. edit distance (reasonably similar)
- Microsoft Bing translations as baseline
- Scripts available to participants for baseline evaluation

Evaluation

- Classical RI evaluation metrics : Precision, Recall and F-measure

- Standard machine translation measure : BLUE (proportion of common n-grams)

- Missing deep analysis of results (they will be ready within one week!)

Outline

Microsoft Bing results

Test		BLUE			Exact			Edit distance		
		1-gr	2-gr	3-gr	P	R	F	P	R	F
en-fr	.25	.51	.24	.17	.99	.23	.38	.63	.63	.63
en-pt	.18	.43	.19	.08	.97	.23	.37	.63	.63	.63
fr-en	.18	.43	.19	.08	.97	.23	.37	.63	.63	.63
fr-pt	.23	.44	.21	.12	.97	.24	.38	.62	.62	.62
pt-en	.17	.50	.19	.05	.97	.24	.38	.56	.56	.56
pt-fr	.22	.48	.20	.15	.97	.23	.37	.51	.51	.51

Cross-lingual Latent Semantic Analysis (CL-LSA)
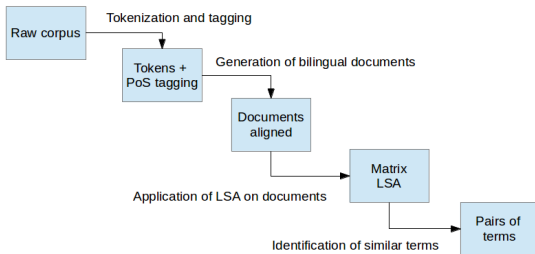
Cross-lingual Latent Semantic Analysis (CL-LSA)

1. Tokenization and tagging
   - Identification of nouns, adjectives and verbs
2. Generation of bilingual documents
   - Concatenation of documents retrieved with the same query in different languages
3. Application of LSA on documents
   - Filter out terms with frequency under 10 times
4. Identification of similar terms
   - Generation of a list containing the top 10 most similar

CL-LSA results

Language Pair	Ontology	Total 1-gram	Found in corpus	First similar	In Top 10
en → fr	cmt	23	21	1	2
	conference	18	17	5	6
	confOf	24	21	5	10
fr → en	cmt	23	21	1	2
	conference	16	16	5	7
	confOf	25	24	7	9
en → pt	cmt	23	19	1	1
	conference	18	17	2	3
	confOf	24	21	7	7
pt → en	cmt	15	15	1	2
	conference	17	16	2	4
	confOf	25	25	8	9
fr → pt	cmt	23	21	2	4
	conference	16	16	4	5
	confOf	25	24	2	5
pt → fr	cmt	15	15	2	3
	conference	17	16	3	4
	confOf	25	25	5	7

- **Found in corpus** : number of 1-gram labels found in corpus
- **First similar** : only the most similar as correct correspondence
- **In Top 10** : correct correspondence if the term is in the top 10 most similar

CL-LSA results

- Some good results to the first positions

Term	Similar
	artigos*
	abreviações
	tabelas
	ortografia
	inglês
paper	nif
	originalidade
	referências
	snif
	infinite

Term	Similar
	autor
	autores
	cigac
	sbz
	rbz
author	flowchart
	zakon
	curr
	editoração
	alçada

*This term is not the exact match (this was an evaluation to verify how good would be to apply LSA in a domain corpus)

CL-LSA results

- But also a lot of bad results

Term	Similar
decision	hidrômetros
	compensação
	matrícula
	aprovação
	anatocismo
	brilhante
	trâmite
	guaxupé
	solicitando
	cedae*

Term	Similar
person	udefa*
	cmdi*
	pcd*
	sydney
	cadastradas
	magicsite
	warszawa
	patos
	ivani
	paulínia

*Such terms are probably acronyms to university or company names. Terms are converted into lowercase during the process.

Outline

Short term perspectives

- Uses Moses for training translation models for the 3 languages
- Apply LSA on Wikipedia articles
- Deep evaluation and comparison of methods (1-gram, 2-gram, 3-gram)
- Provide multiple reference translations for fairer evaluation