Language resources for information extraction and semantic computing

NLP at PUCRS

Renata Vieira, Daniela do Amaral, Lucelene Lopes, Lucas Hilgert, Roger Granada, Sandra Collovini, Evandro Fonseca, Larissa Freitas, Marlo Souza, Artur Freitas, Daniela Schmidt, Bernardo Severo, Cassia Trojahn

Research Themes
Group Mission

- Develop Techniques, Tools and Resources for Natural Language Processing (NLP)
- Generate high specialized human resources for research and practice in NLP

Research Themes
On Going Projects

- Named Entity Recognition
- Term Extraction
- Semantic Relation Identification
- Taxonomic Relations Extraction
- Open Relation Extraction
- Coreference Resolution
- Sentiment Analysis
- Ontology Development
- Ontology Alignment

Named Entity Recognition

Daniela do Amaral

NER consists of the identification and classification of linguistic expressions, mostly proper nouns that refer to a specific entity in the text

We are developing a NER system for portuguese - NERP-CRF

Its first version based on the HAREM corpus and categories.

"A **Universidade de Lisboa** localiza-se em **Portugal**."

Linguamática (2014)

The second version, Geo-NERP-CRF, classifies geological Named Entities (NEs)

"O **Lopingiano** constitui a subdivisão posterior do **Permiano**."

We developed a corpus annotated with 3.500 geological NEs

Resource available at:

- `www.inf.pucrs.br/linatural/NER.html`

Term Extraction
Lucelene Lopes

Term extraction from corpora is the cornerstone of several NLP applications

We developed a software tool for extraction from Portuguese and English Corpora - ExATO

WI (2016)

We developed domain Corpora from areas: Pediatrics, Geology, Data Mining, among others

Resources available at:

- `www.inf.pucrs.br/peg/lucelenelopes/ll_crp.html`

Lists of the extracted terms are available at:

- `www.inf.pucrs.br/peg/lucelenelopes/ll_trm.html`

ENIAC (2013)

Term Extraction

Lucas Hilgert and Lucelene Lopes

We proposed a method to build bilingual dictionaries for specific domanin from parallel corpora

The bilingual dictionaries created are available at:

- `www.inf.pucrs.br/linatural/multilingual`

LREC (2014)

Semantic Relation Identification

Roger Granada

"Words that occur in the same contexts tend to have similar meanings"

Zellig Harris (1954)

English pairs (RG65 [1])			French pairs (JI65 [2])			Portuguese pairs (PT65 [3])		
cord	smile	0.02	corde	sourire	0.00	cordão	sorriso	0.26
rooster	voyage	0.04	coq	périple	0.06	galo	viagem	0.14
noon	string	0.04	midi	ficelle	0.00	almoço	barbante	0.22
fruit	furnace	0.05	fruit	fournaise	0.11	fruta	forno	0.92
autograph	shore	0.06	autographe	rivage	0.00	autógrafo	costa	0.48
automobile	wizard	0.11	automobile	sorcier	0.00	automóvel	bruxo	0.28
mound	stove	0.14	monticule	four	0.06	monte	fogão	0.26
grin	implement	0.18	grimace	instrument	0.00	risada	instrumento	0.70

Semantic Relation Identification

Roger Granada

Generation of a manually annotated dataset for the evaluation of semantic relatedness between word pairs in Portuguese

Lists of word pairs are available at:

- `www.inf.pucrs.br/linatural/wikimodels/similarity.html`

PROPOR (2014)

HREx

- Framework developed in Python
- Methods based on rules – patterns and head-modifier
- Statistical methods – hierarchical clustering, distributional inclusion, document subsumption and entropy

Resource available at:

- `github.com/rogergranada/HREx`

PhD Thesis PUCRS (2015)

Open Relation Extraction

Sandra Collovini

Identifying all possible relations from a text, with no pre-specified definition of the relations

We proposed a process for the extraction of relation descriptors using Conditional Random Fields (CRF)

Relation Triple (NE1, relation descriptor, NE2)

“Ronaldo Lemos **diretor de** Creative Commons.”

Triple (Ronaldo Lemos, **diretor de**, Creative Commons)

IBERAMIA (2014)

Open Relation Extraction

Sandra Collovini

We evaluated a CRF classifier for the extraction of open relation between NEs and pre-defined relation type between these entities

Corpus and relation triples available at:

- `www.inf.pucrs.br/linatural/data_set_RE.html`

LREC (2016)

Coreference Resolution

Evandro Fonseca

Coreference resolution is a process that consists in identifying the several expressions that refer to the same entity in a text:

Após o anúncio de **[o sequenciamento de [o genoma [18]] [56]]** , em a semana passada, **[a França [34]]** resiste como **[único país de [a União_Européia [70]] [34]]** a não permitir **[patenteamento de [genes [26]] [22]]**. **[A UE [70]]** adota, desde junho de 1998 , diretiva favorável a **[o patenteamento de [genes [26]] [22]]**. O texto, redigido por o Parlamento Europeu, Comissão Européia e Conselho de Ministros, utiliza o princípio de que "**[o genoma [18]]** não é patenteável, mas **[a sequência de [um gene [26]] [52]]** pode ser“ no entanto, há restrições. **[O patenteamento [22]]** só pode ser aplicado em pesquisas ligadas a doenças genéticas em que o funcionamento de **[o gene [26]]** é detalhado. **[A França [34]]** é **[o único país [34]]** que se recusa a aceitar a determinação européia . **[A ministra de a Justiça de [o país [34]] , [50]] [Elisabeth Guigou [50]]**, disse que a norma é incompatível com as leis francesas de bioética. Em o início de o mês , **[o CCNE [67]]** (**[Comitê Consultivo Nacional de Ética [67]]**) , **[órgão que orienta o governo francês sobre aspectos éticos de a biotecnologia [67]]** , reforçou a posição de **[a ministra [50]...**

Coreference Resolution

Evandro Fonseca

We introduce semantic knowledge in our coreference model, through Onto-PT

Resource availabe at:

- `http://ontolp.inf.pucrs.br/corref/`

PROPOR (2016)

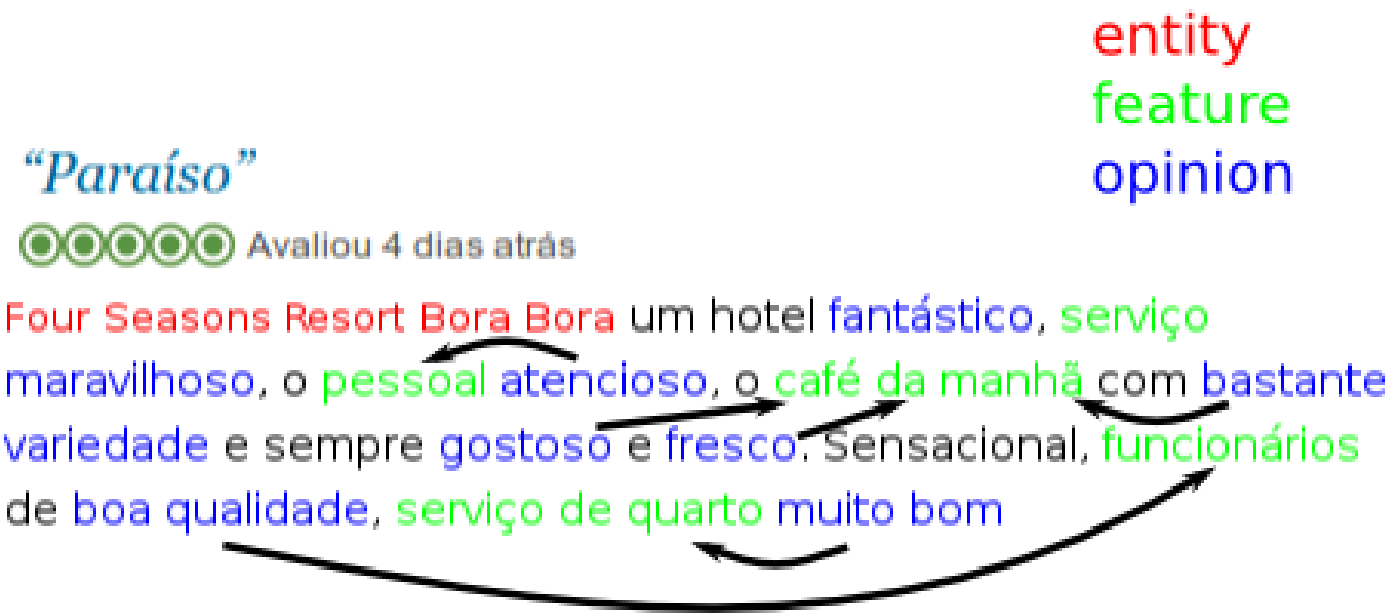
We developed an enriched version of the Summ-it corpus - Summ-it++

Resource available at:

- `www.inf.pucrs.br/linatural/summit_plus.html`

LREC (2016)

Sentiment Analysis is the task of identifying and extracting one's opinions expressed in text



Construction of a Portuguese Opinion Lexicon from multiple resources - OpLexicon

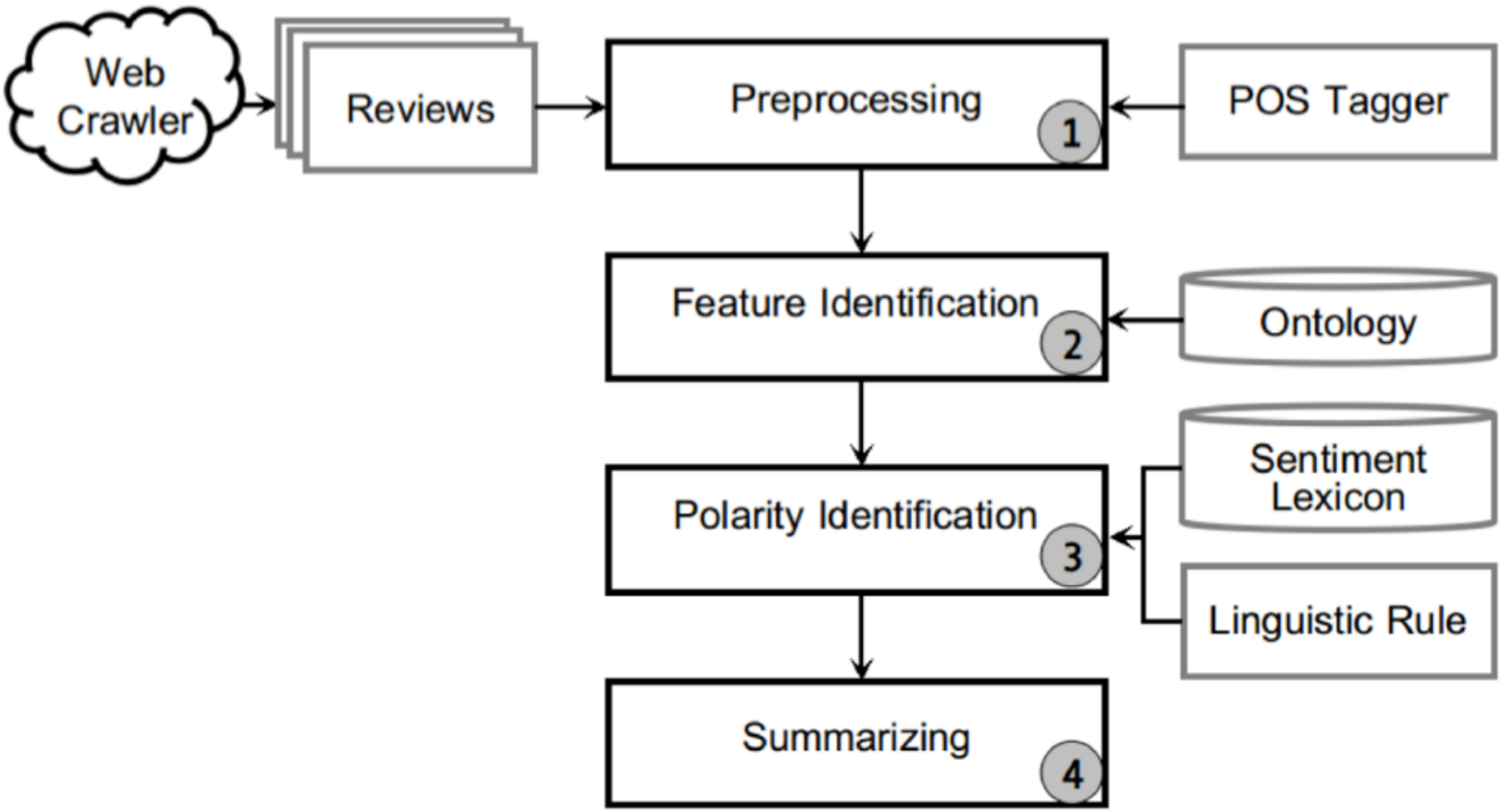
Resource available at:

- `ontolp.inf.pucrs.br/Recursos/downloads-OpLexicon.php`

STIL (2011)

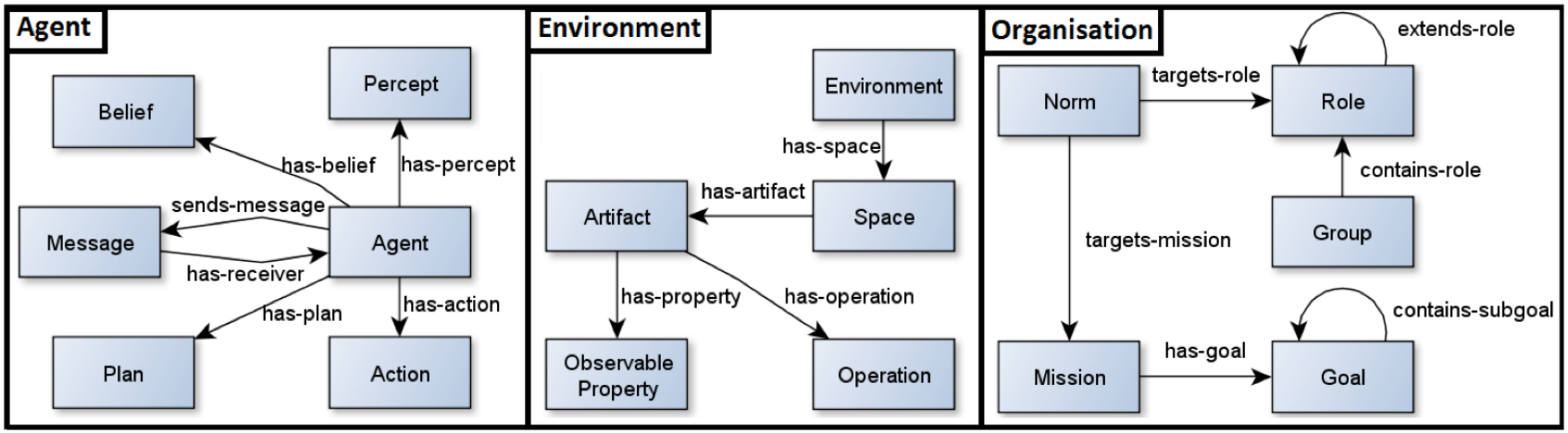
Feature-level sentiment analysis applied to brazilian portuguese reviews

PhD Thesis PUCRS (2015)

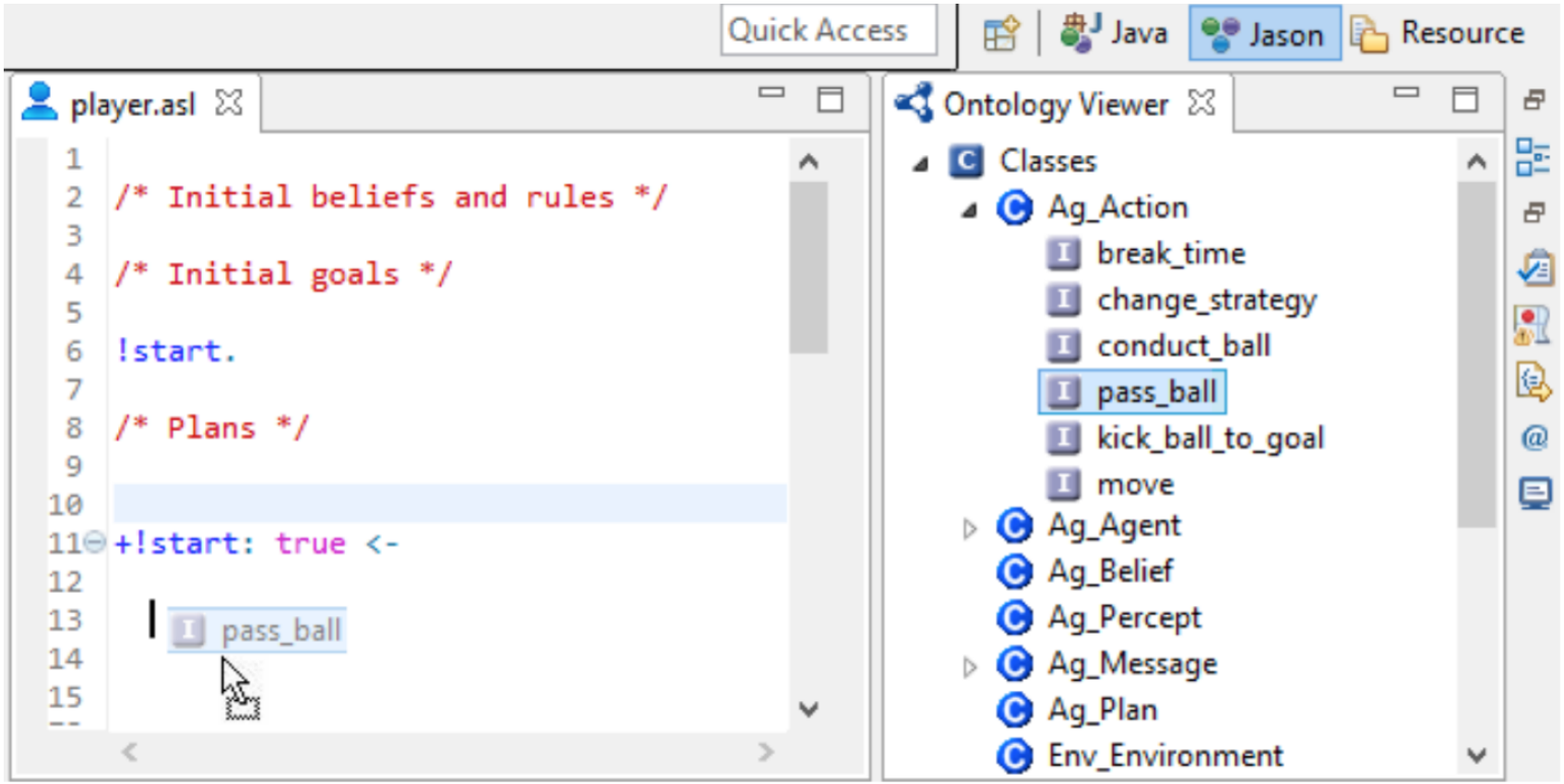


A multi-agent systems engineering tool based on ontologies

ER (2015)



Tool for engineering Multi-Agent Systems using ontology as meta-model



Ontology Alignment

Daniela Schmidt, Bernardo Severo and Cassia Trojahn

Alignment between top-level and domain ontology:

- Analysing the behavior of state-of-the-art matching systems to align different kinds of ontologies (domain and top-level)

Ontology alignment visualization:

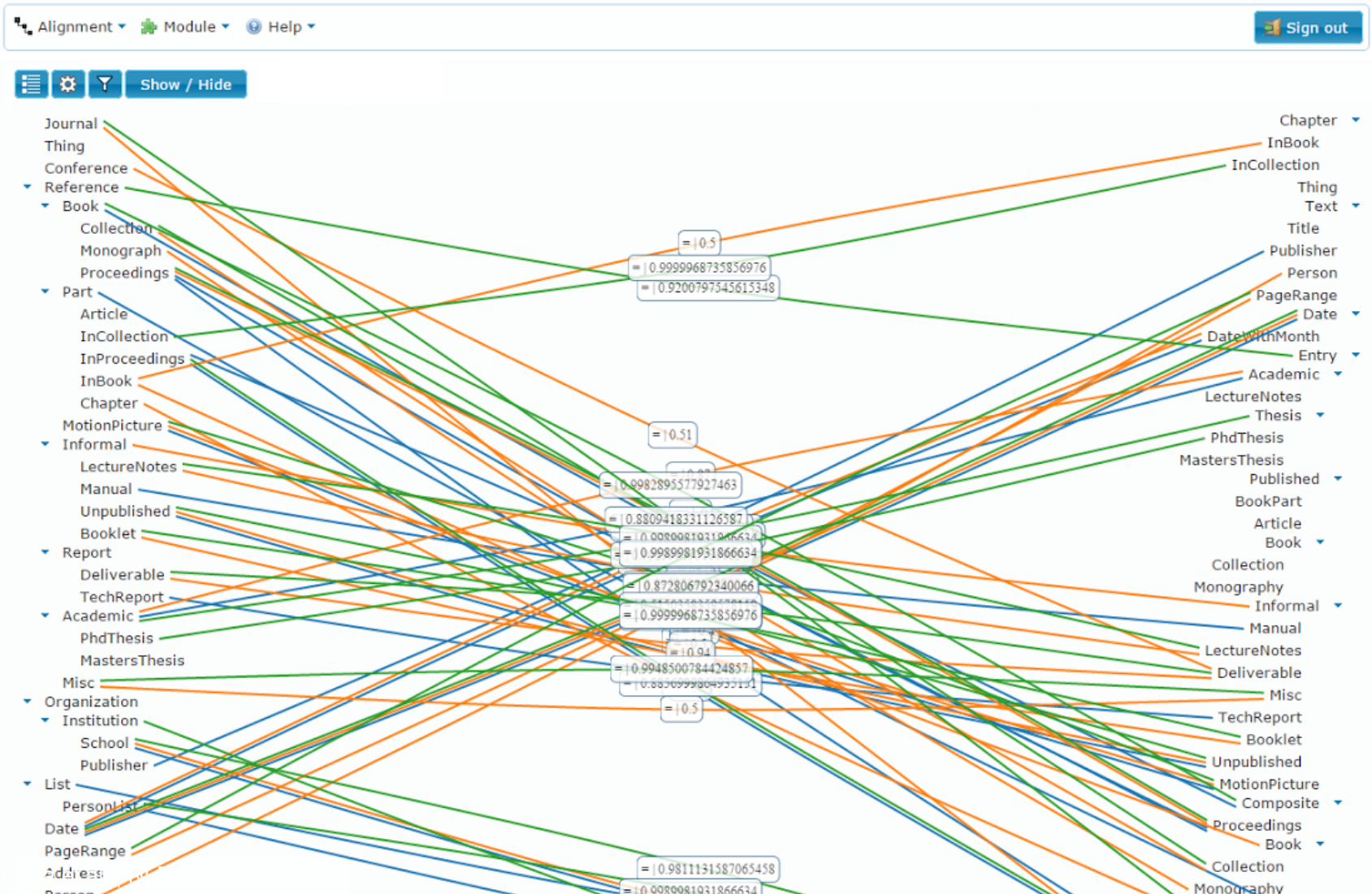
Built an environment for handling ontology alignments with a visual approach - VOAR (Visual Ontology Alignment Environment)

Resource available at:

- `voar.inf.pucrs.br/`

LREC (2014, 2015)

VOAR - Visual Ontology Alignment Environment



Conclusion

- In this paper we presented an overview of currently available language related resources that we have produced at our research lab
- We are happy to share any of the presented research resources with the community


FACIN
PUCRS