



FUNDAÇÃO EDSON QUEIROZ
UNIVERSIDADE DE FORTALEZA
ENSINANDO E APRENDENDO



SIM: A Semantic-Inferentialist Model for Natural Language Processing

Vlória Pinheiro, Knowledge Engineering Laboratory (LEC)
University of Fortaleza
November, 2011
@vladiacelia
vladiacelia@unifor.br

Logics and Ontologies for Portuguese - FGV

Presentation

2

Vlândia Pinheiro, PhD Federal University of Ceará
Professor at **UNIFOR**

UNIFOR

about 25.000 students

Virtual Tour:

<http://www.unifor.br/tourvirtual/>



Knowledge Engineering Laboratory at UNIFOR

3

- ❑ 25 members (2 assistant professor, 1 adjunct professor, 7 MsC, 1 PhD, 9 undergrad students, 5 technical staff) + one startup (Wikinova – www.wikinova.com.br)

- ❑ **Prof. Vasco Furtado**: R&D in the Law Enforcement domain

- ❑ **Prof. Tarcisio Pequeno**: R&D in Logic, Philosophy of Language and Inferentialism

- ❑ Research lines

- ❑ Crowd Mapping

- ❑ Semantic Representation

- ❑ Identification of Malicious activity

- ❑ Reputation and credibility of information

- ❑ Natural Language Processing

- ❑ Multi-agent-based Simulation

Motivação

4


- Modelo teórico abrangente para construção de sistemas eficazes e portáteis que “entendam” termos, sentenças e textos em linguagem natural
 - A quantidade de informações em linguagem natural cresce cada vez mais (jornais na Web, *blogs*, *tweets*, documentos eletrônicos...)
- Sistemas computacionais de entendimento de linguagem natural
 - Sistemas capazes de manipular signos linguísticos para realizar inferências, as quais suportam tomada de decisões, respostas, argumentação, explicação, extração e recuperação de informações etc.

Exemplos Motivadores

5

- Uma mulher, **residente na rua Minas Gerais**, em Piraquara-PR, **foi executada** com um tiro na cabeça, na madrugada de ontem, **pelo amante** Edilson Bezerra Pinto. Os policiais Leandro e Vitor **foram até a rua Santa Catarina e encontraram o corpo da mulher.**
 - *O local do crime foi, provavelmente, a rua Santa Catarina.*
 - *O tipo do crime foi violência doméstica.*
 - *O tipo de arma foi arma de fogo*

 **MUITAS VEZES, A INFORMAÇÃO ESTÁ IMPLÍCITA.**

 **CONHECIMENTO DE MUNDO + CONHECIMENTO LINGUÍSTICO**

Expressão de Conhecimento Semântico

6

Base Semântica	Conhecimento	Relações semânticas
WordNet	Taxonômico	Sinonímia, antonímia, hiponímia/hiperonímia, meronímia/holonímia, similaridade, <i>entailment</i> , causal
FrameNet	Relações entre entidades envolvidas em <i>frames</i>	Específicas por <i>frame</i>
ConceptNet	Senso comum	coisas, espacial, eventual, causal, afetiva, funcional, agente

Problema

7

- ➔ Modelo de expressão semântica representacionista
- Preconiza a expressão de uma representação do mundo
 - Classificação e qualificação semântica de um mundo indutivo
 - Desconsidera os usos dos conceitos



Tarefas de entendimento de linguagem natural

8

Tarefa	Arquitetura	Conhecimento	Raciocínio	Medida-F (inglês)	Medida-F (português)
Anotação de Papéis Semânticos (SRL)	Identificação e classificação de argumentos do verbo (a partir de uma lista pré-especificada de papéis semânticos)	<i>Corpora</i> anotado	Técnicas de Aprendizagem automática Obs: predomínio de atributos sintáticos	73,66%	ndn
Extração de Informação (IE)	Reconhecimento de Entidades Nomeadas (NER)	Léxico próprio, Wikipédia	Regras gramaticais	87%	57,11%
	Extração de relações semânticas entre entidades	Wikipédia, WordNet, <i>Corpora</i> anotado	Técnicas de aprendizagem automática, Regras gramaticais	72,40%	45,02%

Tarefas de entendimento de linguagem natural

9

Tarefa	Arquitetura	Conhecimento	Raciocínio	Medida-F (inglês)	Medida-F (português)
Resposta Automática a Perguntas (QA)	Processamento da pergunta, Recuperação de documentos candidatos, e Seleção de respostas candidatas	WordNet, Corpora anotado	Técnicas de análise sintática, NER, desambiguação de palavras, Lógica Descritiva, Anotação de Papéis Semânticos	70,6%	63,05

Problema

10

- Raciocínio Semântico
 - Processo de “sintatização” do nível semântico
 - Regras gramaticais
 - Regras apreendidas de processamento de *corpus* linguístico
 - Regras de inferência formais
 - Regras *ad hoc*
 - Raciocínio Atomista (de baixo para cima)
- ➔ As inferências são limitadas à informação explícita e desconsideram os usos dos conceitos em situações linguísticas.

Teorias Semânticas Inferencialistas

11

- Wilfrid Sellars (*Inference and Meaning*, 1950)
- Michael Dummett (*Frege's Philosophy of Language*, 1973)
- Robert Brandom (*Articulating Reasons*, 2000)

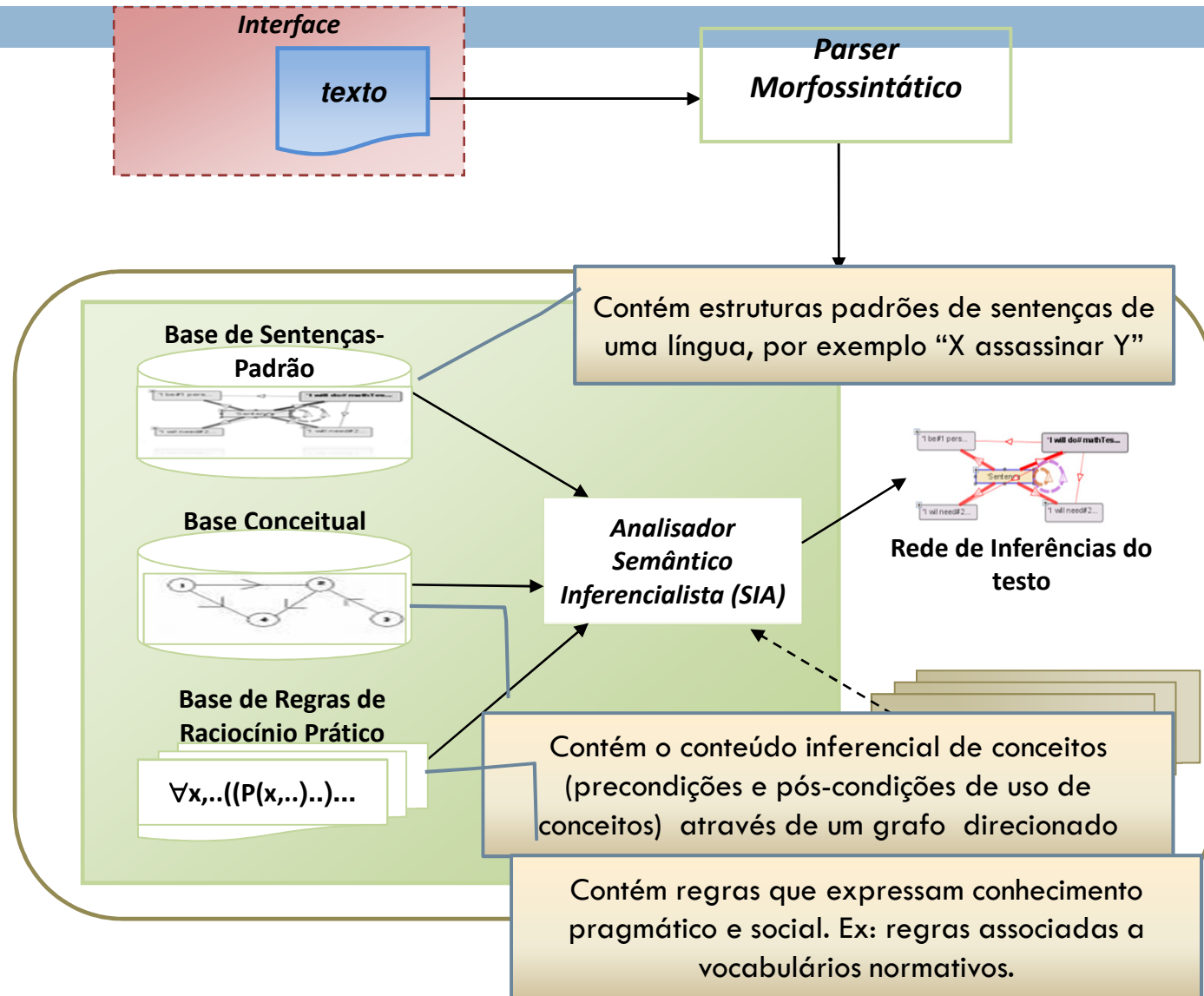
- Entendemos uma sentença quando sabemos defendê-la, argumentar a seu favor, dar explicações, e isto só é possível porque sabemos inferir as **premissas** e **conclusões** de seu proferimento.

- *“to grasp a concept is mastering its inferential use”*

- **O significado de uma sentença em linguagem natural é o conjunto de suas precondições (premissas) e suas pós-condições (conclusões).**

Modelo Semântico Inferencialista - SIM

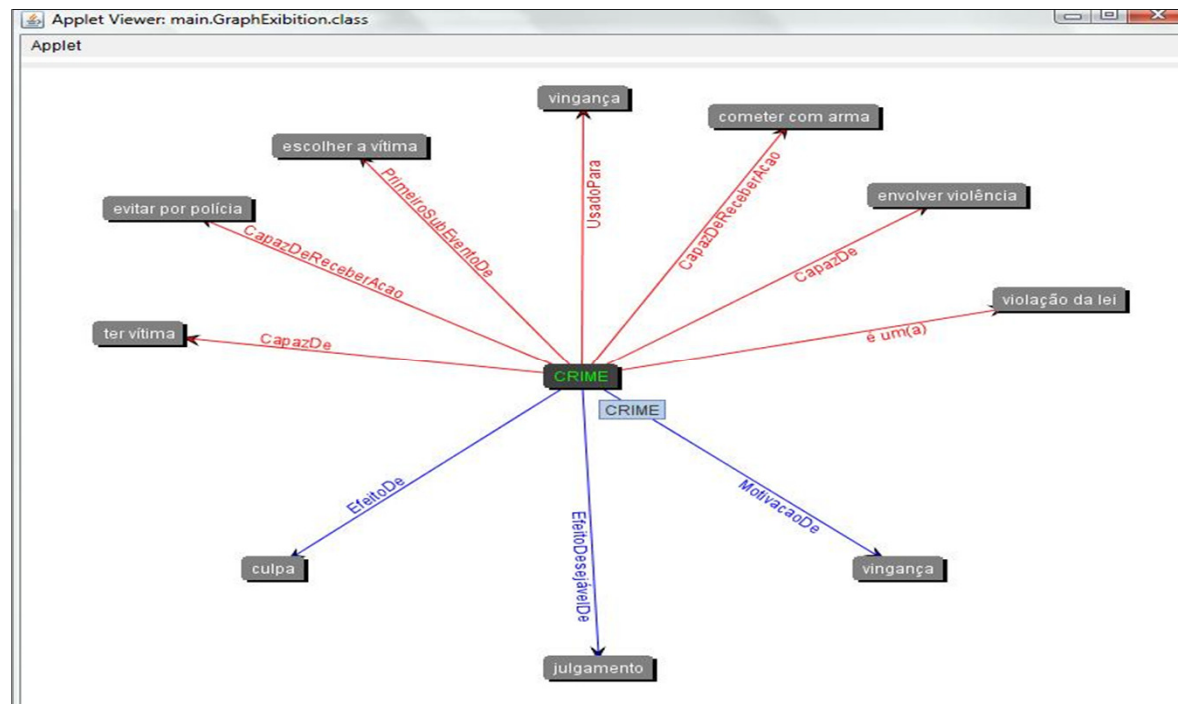
12



SIM — Base Conceitual

13

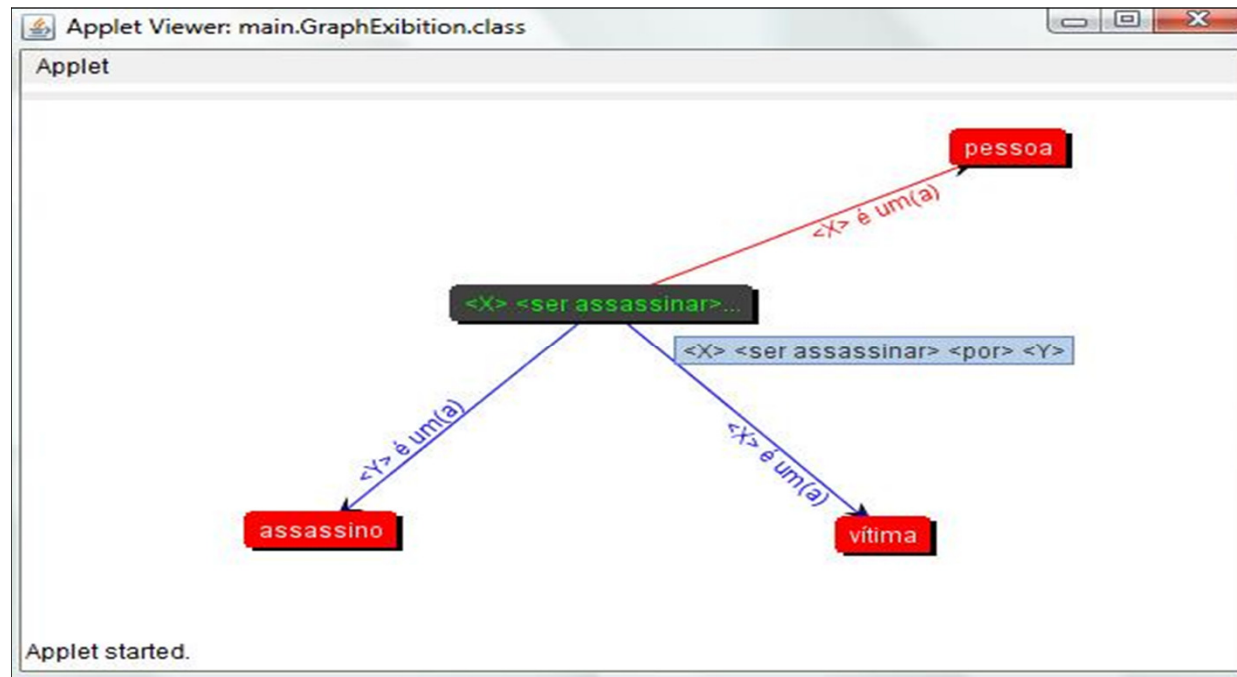
- Precondições e pós-condições de uso dos conceitos
 - Expressas por relações binárias entre dois conceitos:
 - Nome da relação semântica
 - Tipo de relação inferencial: “Pre” ou “Pos”
 - Força da relação inferencial



SIM — Base de Sentenças-Padrão

14

- Precondições e pós-condições de uso das sentenças-padrão
 - Expressas por relações binárias entre um parte da sentença-padrão (nominal, verbal ou complementar) e um conceito da Base Conceitual:
 - Nome da relação semântica
 - Tipo de relação inferencial: “Pre” ou “Pos”



SIA — Analisador Semântico Inferencialista

15

- Raciocínio Material e Holístico
- Medida de Relacionamento Inferencial entre dois conceitos
 - Desambiguar termos homônimos
 - Definir a contribuição semântica dos conceitos
- 03 (três) formas de raciocínio semântico para geração de premissas e conclusões da sentença **s**

SIA — Medida de Relacionamento Inferencial

16

Quanto mais as circunstâncias e conseqüências de uso de dois conceitos são semelhantes mais eles podem ser usados em fluxos de raciocínio semelhantes

$$\theta_{c_1,c_2} = (F_1\omega_1 + F_2\omega_2 + F_3\omega_3)\mu_{c_1,c_2}$$

- F_1, F_2, F_3 são os somatórios das forças das relações inferenciais de c_1 e c_2 que satisfazem a três formas de proximidades inferenciais,
- $\omega_1, \omega_2, \omega_3$ são os pesos, atribuídos por parâmetro, das três formas de proximidades inferenciais, e
- μ_{c_1,c_2} é o fator de normalização entre os conceitos c_1 e c_2 .

SIA — Raciocínio Semântico Inferencialista

17

1. Geração de premissas e conclusões da sentença s com base no conteúdo inferencial de conceitos c_i usados em s

$$\frac{(nome_relacao, c_1, c_2, "Pre"), s(c_1)}{("Pre", s(c_1), s(c_1|c_2))} (E_1 - c)$$

Exemplo:

Sejam

- s_1 = "O crime ocorreu na Rua Titan, 33"

- c_1 = "crime" = $nucleo(sn(s_1))$

- pré-condição de c_1 : (*éUm*, 'crime', 'violação da lei', 'Pre')

Logo, por (E₁-c), pode ser gerada a relação PreCondicacao (s_1, s_2), onde s_2 = "<Um(a)> <violação da lei> <ocorreu> <na Rua Titan, 33>"

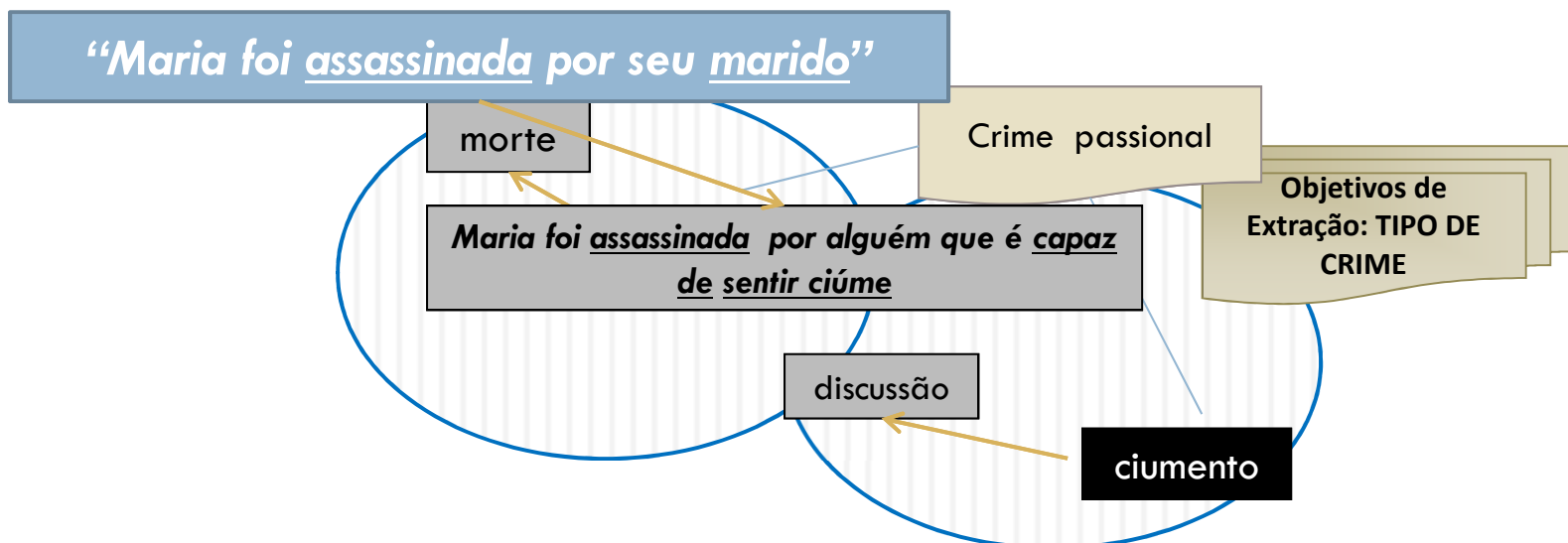
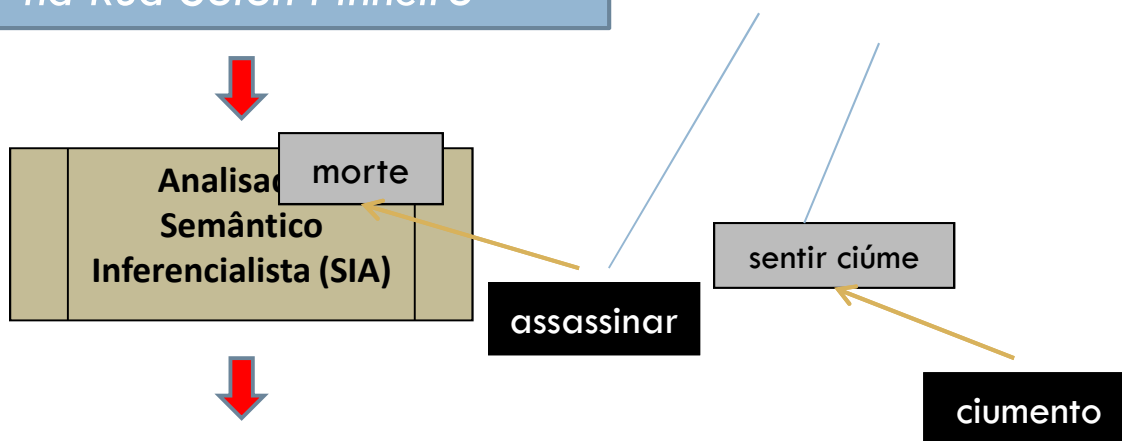
SIA – Exemplo

18

Maria foi assassinada por seu marido depois de uma discussão na Rua Solon Pinheiro

Legenda:

- ➡ Processo
- ➡ Pós-condição
- Associação



SIM — Características

19

- Conteúdo semântico que expressa situações de uso de conceitos e sentenças
 - Mecanismo de raciocínio material e holístico
- ➔ Arcabouço inferencial que considera o aspecto pragmático da linguagem.

Construindo a Base Conceitual

20

Geração de conteúdo inferencial a partir da ConceptNet.Tr

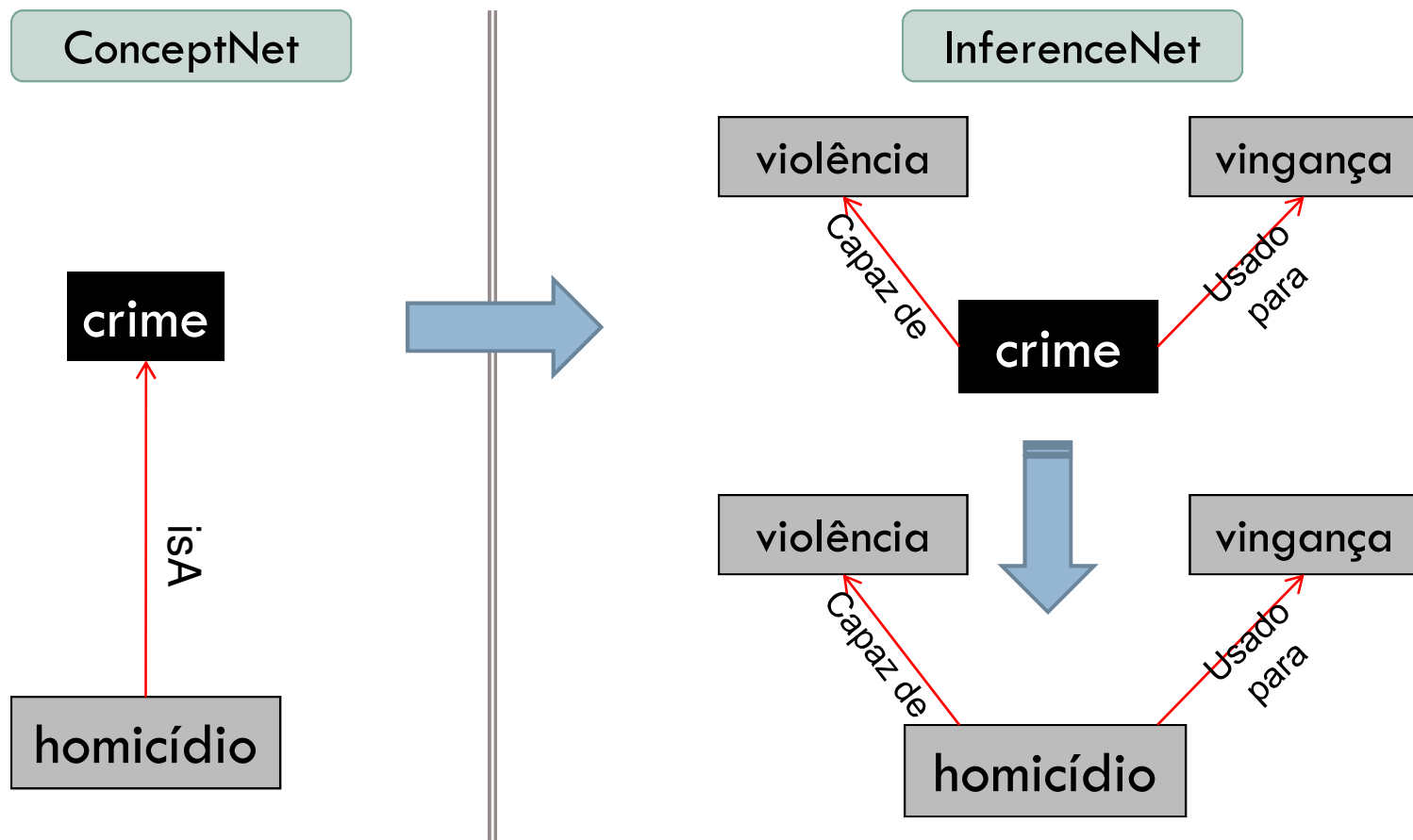
- Cada tipo de relação da ConceptNet.Tr $tipo_rel(c_1, c_2)$ \rightarrow uma pré-condição ou uma pós-condição de uso do conceito c_1

Categoria	Tipo de Relação (ConceptNet)	Tipo de Relação Inferencial (InferenceNet.BR)
EVENTOS	PrerequisiteEventOf; FirstSubeventOf; SubeventOf; LastSubeventOf	pré-condição
CAUSAL	EffectOf; DesirousEffectOf	pós-condição

Construindo a Base Conceitual

21

Geração de precondições [pós-condição] a partir de relações de especialização *IsA* ou *DefinedAs*



Extração de Verbo+Preposição do corpus CRIMES2008 – notícias de crimes publicadas em jornais do Brasil em 2008 contendo 150k palavras (4 meses de notícias)

- “*Segundo a Polícia, dois homens que ocupavam uma moto assassinaram o bancário*”
- *dois homens assassinaram o bancário*”
- “*<X> <assassinar> <Y>*”

Geração de conteúdo inferencial de sentenças-padrão

- Pós-condições de acordo com a circunstância expressa pelo complemento adverbial (lugar, tempo e causa)
 - Exemplo
 - “<X> <assassinar> <em frente de> <Y>”
 - → pós-condição: *ehUm (Y, "local")*
- Precondições [pós-condições] relacionadas ao autor e a vítima de crimes para verbos “semanticamente relacionados” a crime (usando a Medida de Relacionamento Inferencial)
 - Exemplo
 - “<X> <assassinar> <Y>”
 - → pós-condição: *ehUm (Y, "vitima")*
 - → precondição: *ehUm (X, "pessoa")*
 - → pós-condição: *ehUm (X, "assassino")*

Números da InferenceNet.BR 1.0

24

Elementos da Base	InferenceNet.BR	ConceptNet 2.1	WordNet 3.0	FrameNet II
BASE CONCEITUAL				
Conceitos	182.170	182.162	117.659	11.836
Relações entre conceitos	674.857	1,6 milhão	s/informação	-
- precondições	620.851	-	-	-
- pós-condições	54.006	-	-	-
BASE DE SENTENÇAS-PADRÃO				
Sentenças-Padrão	5.910	-	-	969
Relações entre sentenças-padrão	1.432	-	-	s/informação
- precondições	328	-	-	-
- pós-condições	1.104	-	-	-

wikicrimes.org

25



Mapeando crimes colaborativamente

vasco@unifor.br | [Logout](#) | [Minha Reputacao](#) | [Minhas Áreas](#) | [Meus Registros](#) | [Minha Conta](#)

Compartilhe informações sobre crimes

Salva onde não é seguro!

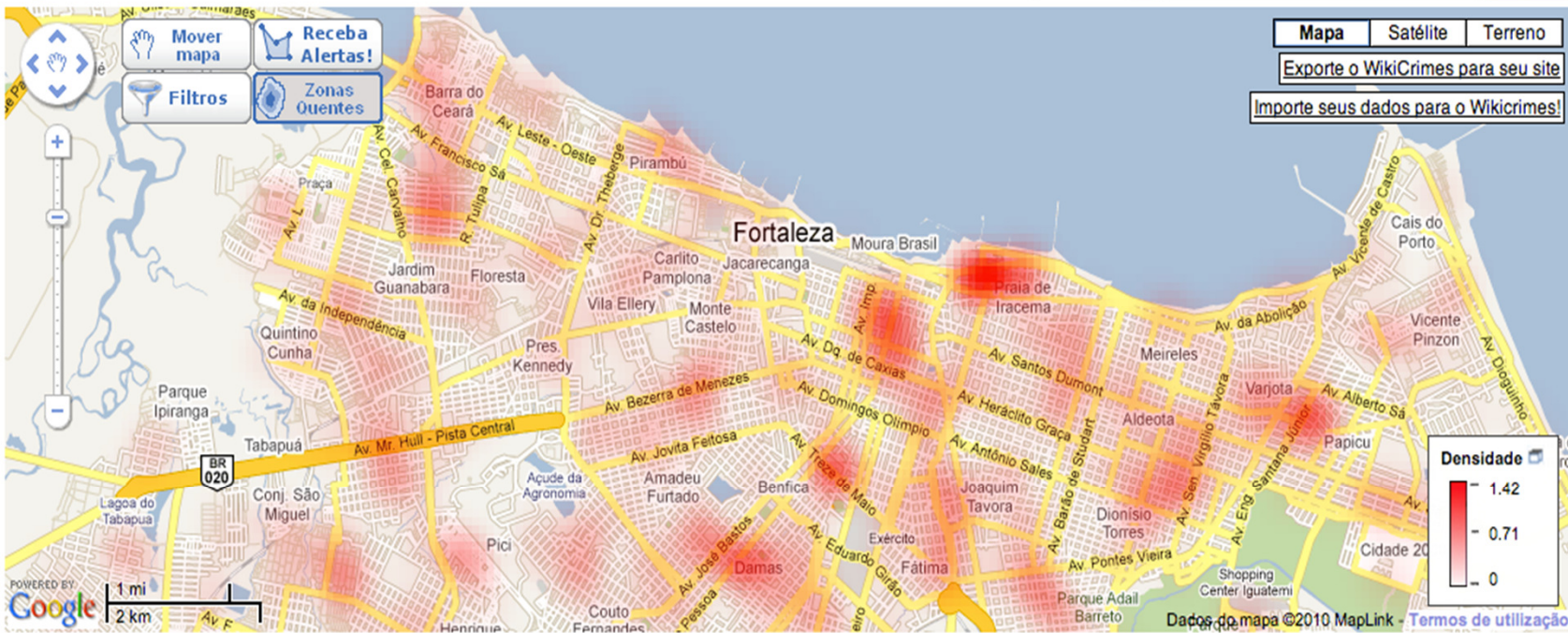
13121 Crimes Registrados

[Registre Crime](#)

[Denuncie tráfico/violência](#)

Procurar...

[pesquisar](#)



Informações da área visível do mapa:

Total de ocorrências: 737

Filtros aplicados: Credibilidade, Data

[Mais estatísticas!](#)

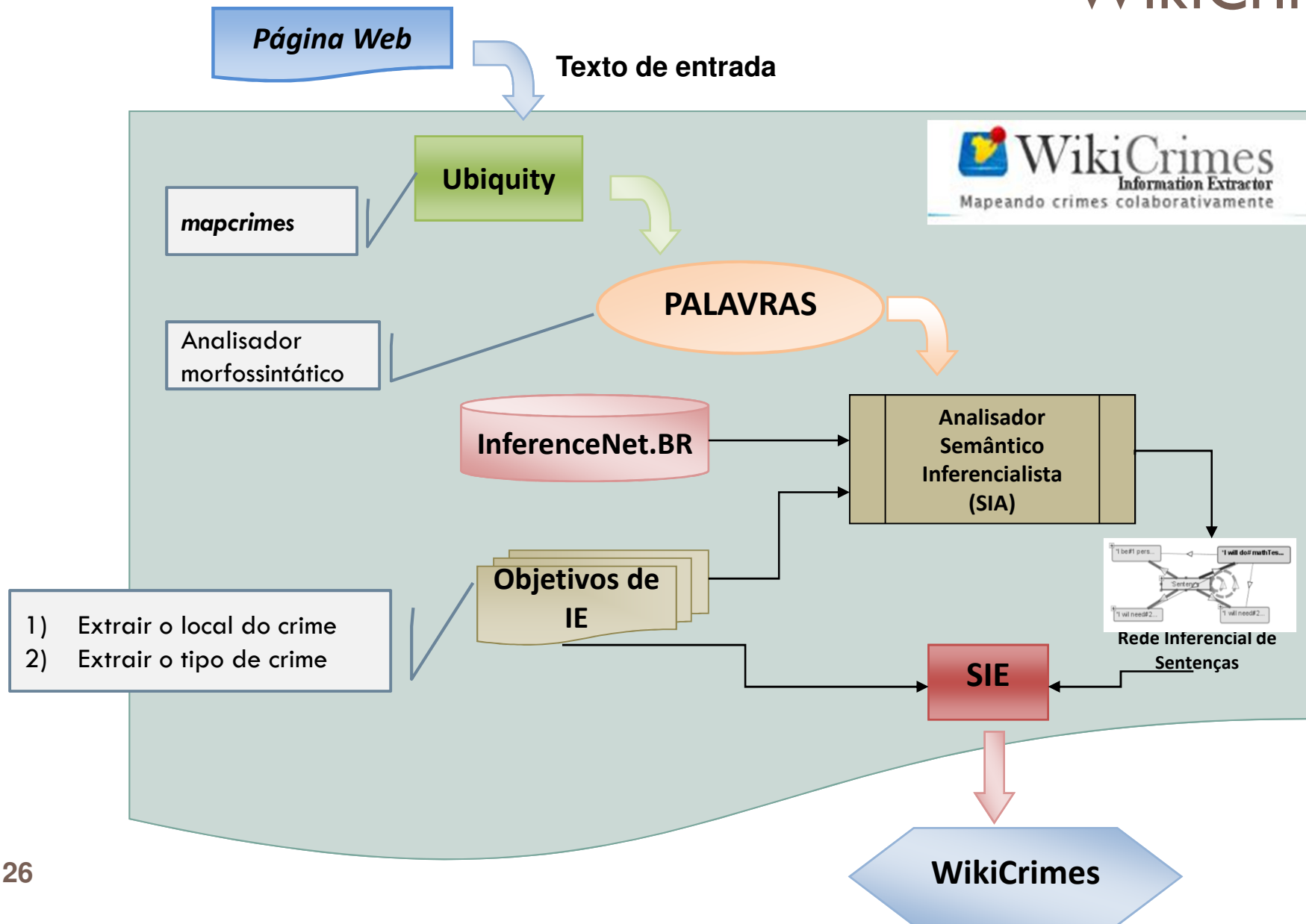


- Robo (38.40%)
- Furto (24.69%)
- Denuncia (2.44%)
- Outro (34.46%)



- Falta de Policiamento (28.07%)
- Crime organizado (15.79%)
- Rota de fuga fácil (11.40%)
- Outros (44.74%)

Arquitetura do Extrator de Informações para WikiCrimes



Textos avaliacao WikicrimesIE - Mozilla Firefox

Arquivo Editar Exibir Histórico Favoritos Ferramentas Ajuda

http://www.surveymonkey.com/s.aspx?sm=yfz3i4W04fZH_2bz_2fUHNrP5A_3d_3d

Mais visitados Últimas notícias Questionário sobre a f... Textos avaliacao Wiki... Copy of Textos avalia... ubiquity-firefox-0.1.6... http://localhost:8080/...

FoxLingo Página web Texto Serviços AutoTrans Pesquisar

http://192.168.0.9:...E/wikicrimesIE.html Bem-Vindo | WikiCrimes.org Textos avaliacao WikicrimesIE

mapc

mapcrimes
MAIS UM CRIME COM CARACTERÍSTICAS ... (in text)

mapcrimes (text) in
MAIS UM CRIME COM CARACTERÍSTICAS ...

Creates a Crime Ocurrence in with these contents: Mais um crime com características de execução sumária foi registrado em Fortaleza. Na noite de terça-feira, o jovem Marcelo dos Santos Vasconcelos, 29, foi fuzilado na porta de casa. O crime ocorreu na Rua Casimiro de Abreu, em Parangaba.

Exit this survey

Marcelo dos Santos

Mais um crime com características de execução sumária foi registrado em Fortaleza. Na noite de terça-feira, o jovem Marcelo dos Santos Vasconcelos, 29, foi fuzilado na porta de casa. O crime ocorreu na Rua Casimiro de Abreu, em Parangaba

zotero

Iniciar Java - Eclipse Platform Textos avaliacao Wiki... imagemWikiCrimesIE... 16:20

Dados Sobre Endereço

Endereço

R. Casimiro de Abreu

Cidade

Fortaleza

Estado

CE

País

BR

CEP

60710-250

Tipo do Local (*)

Selecione...

Dados Sobre o Crime

Data da Ocorrência (dd/mm/yyyy) (*)

Horário (*)

Selecione...

Quantidade de Criminosos

Quantidade de Vítimas

Qual a sua relação com o crime? (*)

Selecione...

A polícia foi informada? (*)

Selecione...

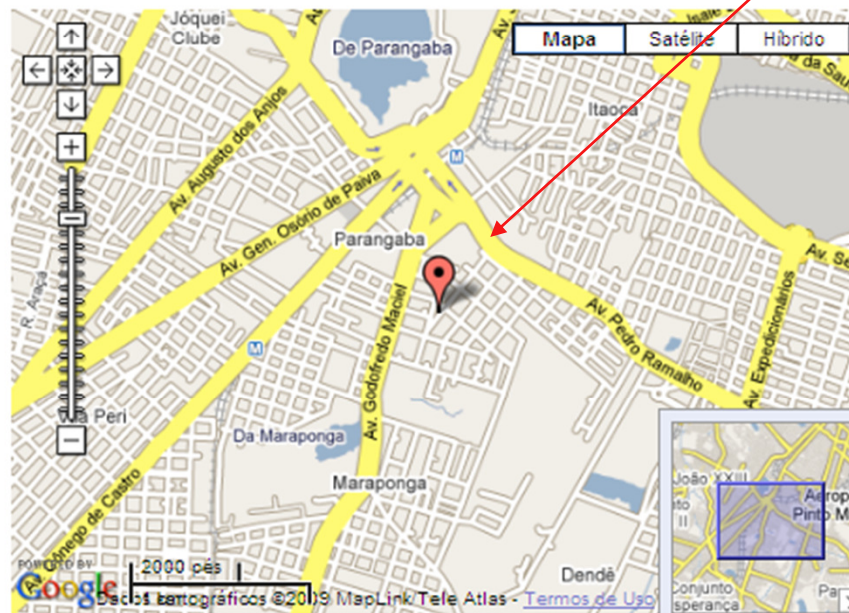
Texto Selecionado

Texto Selecionado pelo [Ubiquity](#). Descrição: (*)

Mais um crime com características de execução sumária foi registrado em Fortaleza. Na noite de terça-feira, o jovem Marcelo dos Santos Vasconcelos, 29, foi fuzilado na porta de casa. O crime ocorreu na Rua Casimiro de Abreu, em Parangaba. Crime encontrado na pagina [http://www.surveymonkey.com/s.aspx?sm=yfz3i4W04fZH\\$2bz\\$2fUHNrP5A\\$3d\\$3d](http://www.surveymonkey.com/s.aspx?sm=yfz3i4W04fZH$2bz$2fUHNrP5A$3d$3d)

Endereço obtido pelo SIA: Rua Casimiro de Abre

Mapa



Pesquisar no Mapa

Dados sobre o Crime

Tipo do Crime (*)

Homicídio

Tipo de Vítima (*)

Selecione...

Arma Utilizada (*)

Selecione...

O que você considera causa/motivo desta ocorrência. Marque no mínimo 1 e no máximo 4 opções:

- | | |
|---|---|
| <input type="checkbox"/> Má iluminação pública | <input type="checkbox"/> Ausência de lazer para os jovens |
| <input type="checkbox"/> Desemprego na região | <input type="checkbox"/> Rota de fuga fácil |
| <input type="checkbox"/> Disputa de gangues | <input type="checkbox"/> Uso/Tráfico de drogas |
| <input type="checkbox"/> Uso de álcool | <input type="checkbox"/> Crianças/Adolescentes nas ruas |
| <input type="checkbox"/> Alta concentração de pessoas | <input type="checkbox"/> Falta de Policiamento |
| <input type="checkbox"/> Omissão das testemunhas | <input type="checkbox"/> Proximidade de regiões perigosas |
| <input type="checkbox"/> Impunidade Penal | <input type="checkbox"/> Pistolagem |
| <input type="checkbox"/> Violência Policial | <input type="checkbox"/> Falta de Moradia/Má urbanização |
| <input type="checkbox"/> Crime organizado | <input type="checkbox"/> Crime passional |
| <input checked="" type="checkbox"/> Outros | <input type="checkbox"/> Não Sei |

(*) Campos Obrigatórios

Avaliação do SIM

29

- Coleção Dourada com 200 crimes anotados com respostas de dois especialistas

Medida	Local de Crime	Tipo de Crime	Causa do Crime	Tipo de Arma	Média
Precisão	87%	72%	76%	85%	80%
Cobertura	71%	68%	70%	76%	71%
Medida-F	78%	70%	73%	80%	75%
Erros de análise sintática	2%	7%	7%	7%	-

Avaliação Quantitativa do SIM

30

- SIM:
 - Medida-F = **75%** (português – local e tipo de crime)
- Melhor sistema na tarefa de SRL no CoNLL-2009:
 - Medida-F = **73,66%** (inglês)
- Melhor sistema na tarefa de NER no 2º.HAREM-2008 (na categoria LOCAL):
 - Medida-F = **59,93%** (português)
- **ATENÇÃO !!!**
 - Informações explícitas
 - Dependência de *corpora* anotados
 - Dependência de analisador morfossintático

Contribuições da Pesquisa

31

- Novo modelo para expressão e raciocínio semântico de linguagem natural – o Modelo Semântico Inferencialista (SIM).
 - Semântica Computacional Inferencialista
- O primeiro recurso linguístico com um conteúdo inferencialista para a língua portuguesa – InferenceNet.BR – contendo em torno de 190.000 conceitos, 700.000 relações inferenciais entre conceitos, 6000 sentenças-padrão e 1500 relações inferenciais de sentenças-padrão.
- Um componente de software que implementa o algoritmo SIA, o qual pode ser reusado em diversas aplicações de PLN.

Contribuições da Pesquisa

32

- Uma medida de relacionamento semântico que pode ser usada em diversas aplicações e tarefas de PLN.
 - Resolução de anáforas (Dissertação Mestrado UFC em 2010)
- O portal www.inferencenet.org contendo serviços para a comunidade de PLN, que permitem a consulta, evolução e disseminação da base InferenceNet.BR.
- O Extrator de Informações para o sistema WikiCrimes — WikiCrimesIE.
 - Componentes genéricos para sistemas de extração de informações

Trabalhos em andamento e futuros

33

- Em andamento
 - Melhorias de engenharia do SIM e da InferenceNet.BR
 - InferenceNet for LOD cloud
 - Avaliação intrínseca do recurso InferenceNet.BR
 - Aprendizagem de conhecimento Inferencialista
 - Uso do modelo em outras tarefas de PLN: semantic web annotation
- Futuros
 - Evolução do algoritmo SIA para revisão e atualização da rede inferencial no decorrer de uma situação linguística
 - Novos mecanismos de inferência para combinar conteúdo inferencial de conceitos e sentenças
 - Raciocínio holístico sobre texto e contexto

Referências

34

- www.inferencenet.org
- Pinheiro, V., Pequeno, T., Furtado, V., Franco, W. InferenceNet.Br: Expression of Inferentialist Semantic Content of the Portuguese Language. PROPOR 2010.
- Pinheiro, V., Pequeno, T., Furtado, V., Nogueira, D. *Natural Language Processing Based on Semantic Inferentialism for Extracting Crime Information from Text*. IEEE ISI 2010. **Best Paper Award**
- Pinheiro, V., Pequeno, T., Furtado, V. Um Analisador Semântico Inferencialista de Sentenças em Linguagem Natural. Linguamática 2010
- Pinheiro, V., Pequeno, T., Furtado, V., Nogueira, D. Information Extraction from Text Based on Semantic Inferentialism. FQAS 2009.



Obrigada !

vladiacelia@unifor.br