Processamento de Linguagem Natural em textos da História Comptemporânea do Brasil: o projeto OpenWordnet-PT

Alexandre Rademaker

EMAp, FGV

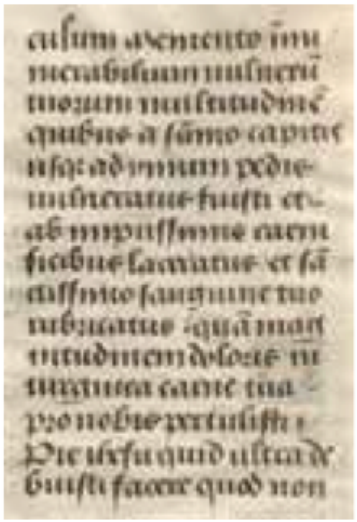
October 3, 2012

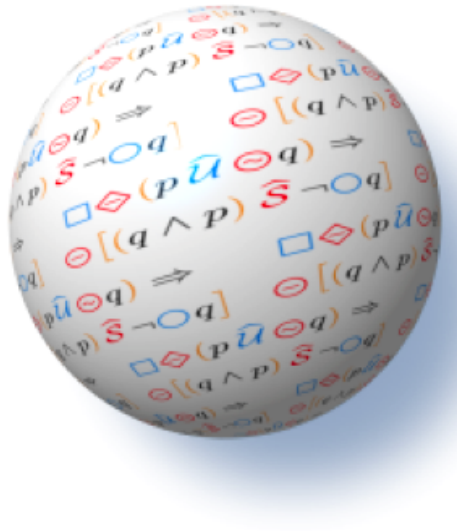
Colaboradores

- Valeria de Paiva
- Gerard de Melo, Berkeley
- Adam Pease, `http://www.articulatesoftware.com`
- Rafael Haeusler
- E outros.

Conteúdo

Processamento de linguagem natural para um Lógico





Thanks Valeria de Paiva.

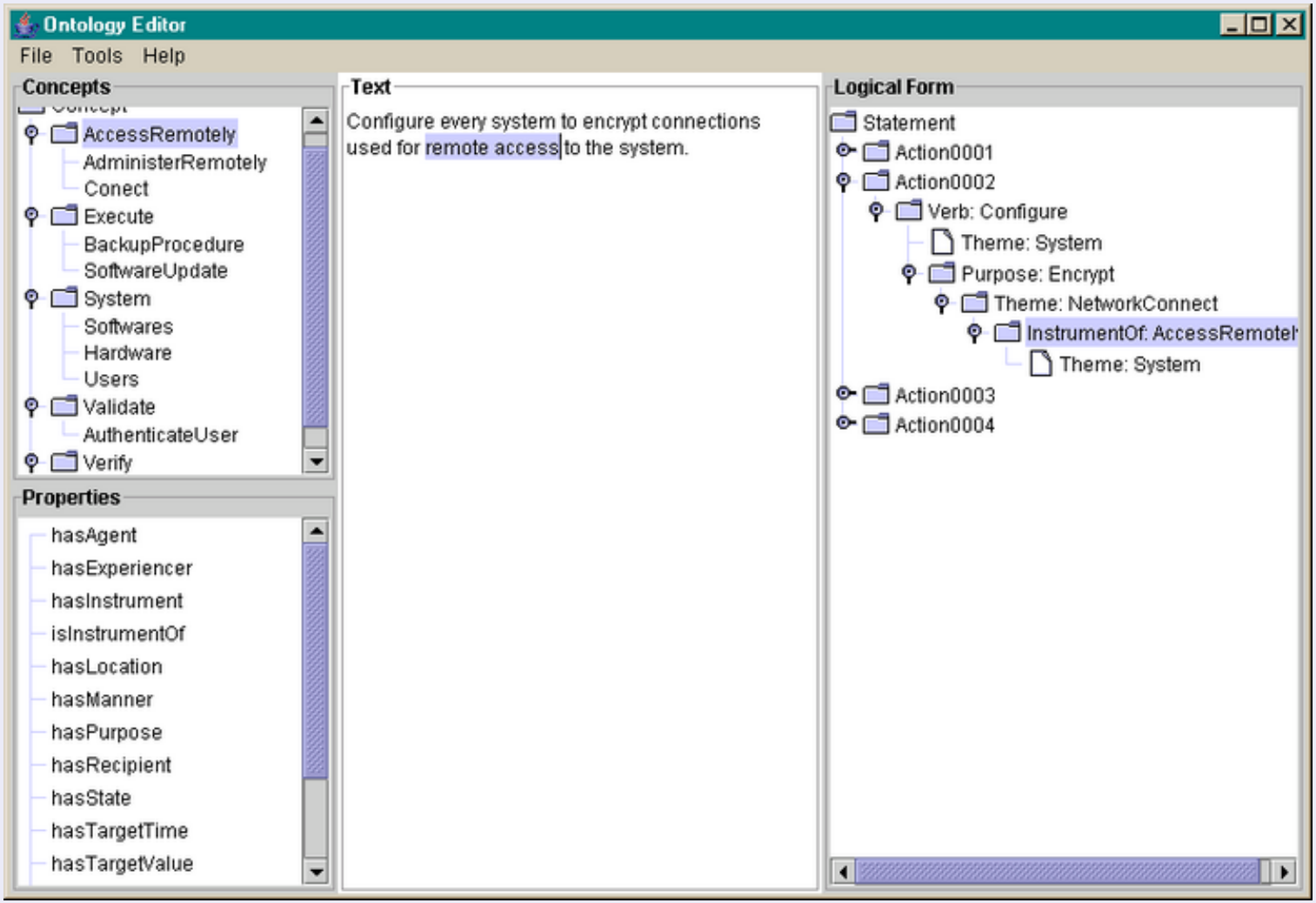
Experiência prévia: projeto ANUBIS

Configure every system to encrypt connections used for remote access to the system.

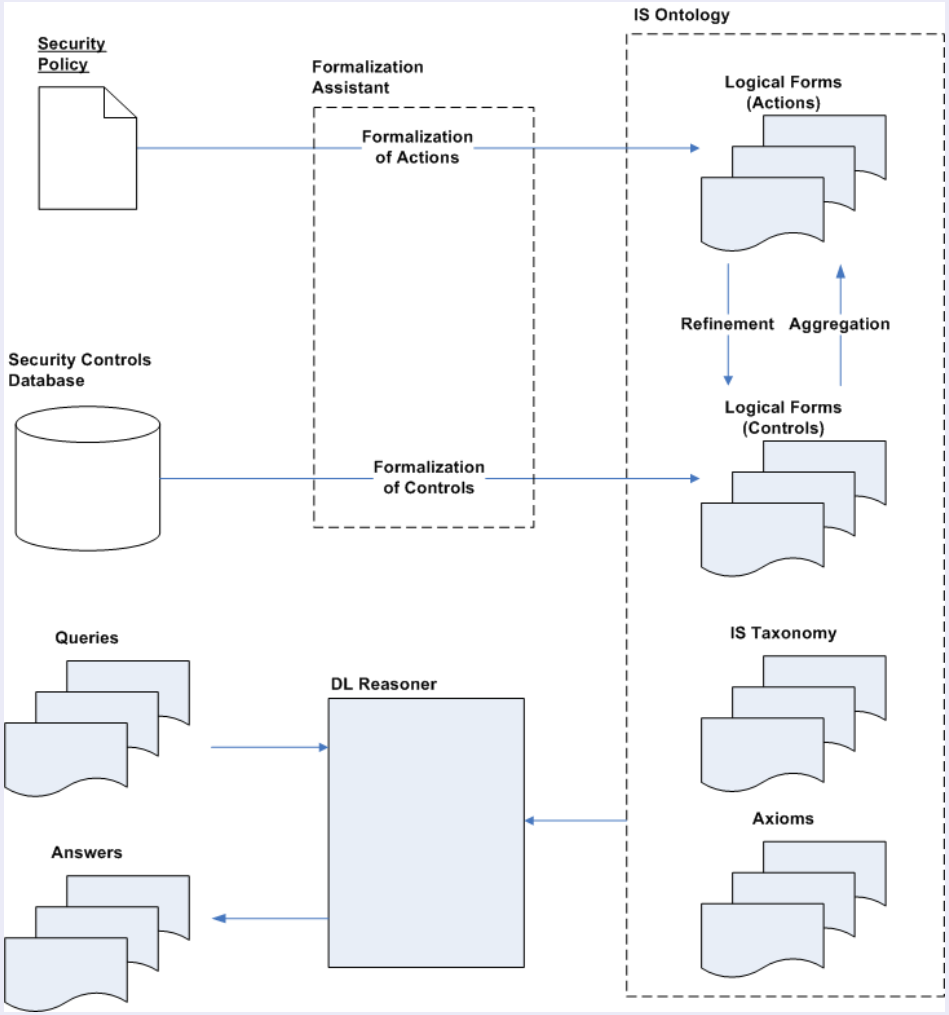
Representação lógica (logical forms):

$$\texttt{Action01} \equiv \exists \texttt{hasVerb}.(\mathit{Configure} \sqcap \exists \mathit{hasTheme}.\mathit{System} \sqcap \exists \mathit{hasPurpose}.(\mathit{Encrypt} \sqcap \exists \mathit{hasTheme}.(\mathit{NetworkConnect} \sqcap \exists \mathit{isInstrumentOf}.(\mathit{AccessRemotely} \sqcap \exists \mathit{hasTheme}.\mathit{System}))))$$

Experiência prévia: projeto ANUBIS



Experiência prévia: projeto ANUBIS



Usando a lógica *iALC* para formalização de leis

Um exemplo

Peter and Maria signed a renting contract. The subject of the contract is an apartment in Rio de Janeiro. The contract states that any dispute will go to court in Rio de Janeiro. Peter is 17 and Maria is 20. Peter lives in Edinburgh and Maria lives in Rio.

Only legally capable individuals have civil obligations:
$PeterLiable \preceq ContractHolds@RioCourt$, shortly, $pl \preceq cmp$

$MariaLiable \preceq ContractHolds@RioCourt$, shortly, $ml \preceq cmp$

Concepts, nominals and their relationships
BR is the collection of Brazilian Valid Legal Statements
SC is the collection of Scottish Valid Legal Statements
PIL_{BR} is the collection of Private International Laws in Brazil
$ABROAD$ is the collection of VLS outside Brazil
$LexDomicilium$ is a legal connection:

Legal Connections The pair $\langle pl, pl \rangle$ is in $LexDomicilium$

Axiomas não lógicos

O conjunto Δ, de conceitos, e Ω, de *iALC* axiomas representam o conhecimento extraido do caso.

$$\Delta = \boxed{\begin{array}{ccc} ml : \text{BR} & pl : \text{SC} & pl \preceq cmp \\ ml \preceq cmp & pl \text{ LexDom } pl & \end{array}}$$

$$\Omega = \boxed{\begin{array}{c} \text{PIL}_{\text{BR}} \Rightarrow \text{BR} \\ \text{SC} \Rightarrow \text{ABROAD} \\ \exists \text{LexD}_1.\text{L}_1 \ldots \sqcup \exists \text{LexDom.ABROAD} \sqcup \ldots \exists \text{LexD}_k.\text{L}_k \Rightarrow \text{PIL}_{\text{BR}} \end{array}}$$

Um sistema dedutivo para *iALC*

Usual Structural-Rules for Intuitionistic Logic

$$\frac{}{\Gamma, x\colon C \Rightarrow x\colon C, \Delta}$$

$$\frac{\Gamma_1 \Rightarrow C \qquad \Gamma_2, D \Rightarrow \Delta}{\Gamma_1, \Gamma_2, C \sqsubseteq D \Rightarrow \Delta} \sqsubseteq\text{-l}$$

$$\frac{\Gamma, x\colon C, x\colon D \Rightarrow \Delta}{\Gamma, x\colon (C \sqcap D) \Rightarrow \Delta} \sqcap\text{-l}$$

$$\frac{\Gamma, x\colon C \Rightarrow \Delta \qquad \Gamma, x\colon D \Rightarrow \Delta}{\Gamma, x\colon (C \sqcup D), \Rightarrow \Delta} \sqcup\text{-l}$$

$$\frac{\Gamma, x\colon \forall R.C, y\colon C, xRy \Rightarrow \Delta}{\Gamma, x\colon \forall R.C, xRy \Rightarrow \Delta} \forall\text{-l}$$

$$\frac{\Gamma, xRy, y\colon C \Rightarrow \Delta}{\Gamma, x\colon \exists R.C \Rightarrow \Delta} \exists\text{-l}$$

$$\frac{\Delta \Rightarrow x : A \qquad A \Rightarrow B}{\Delta \Rightarrow x : B} \in\text{-r}$$

$$\frac{}{xRy, \Gamma \Rightarrow \Delta, xRy}$$

$$\frac{\Gamma, C \Rightarrow D}{\Gamma \Rightarrow C \sqsubseteq D} \sqsubseteq\text{-r}$$

$$\frac{\Gamma \Rightarrow x\colon C, \Delta \qquad \Gamma \Rightarrow x\colon D, \Delta}{\Gamma \Rightarrow x\colon (C \sqcap D), \Delta} \sqcap\text{-r}$$

$$\frac{\Gamma \Rightarrow x\colon C, x\colon D, \Delta}{\Gamma \Rightarrow x\colon (C \sqcup D), \Delta}$$

$$\frac{\Gamma, xRy \Rightarrow y\colon C, \Delta}{\Gamma \Rightarrow x\colon \forall R.C, \Delta} \forall\text{-r}$$

$$\frac{\Gamma \Rightarrow \Delta, xRy \qquad \Gamma \Rightarrow \Delta, y\colon C}{\Gamma \Rightarrow \Delta, x\colon \exists R.C} \exists\text{-r}$$

Usando o sistema dedutivo

$$\dfrac{\dfrac{\Delta \Rightarrow pl : \mathrm{SC} \quad \dfrac{\Omega}{pl : \mathrm{SC} \Rightarrow pl : \mathrm{A}}}{\Delta \Rightarrow pl : \mathrm{A}}\,cut \quad \Delta \Rightarrow pl\ \mathrm{LexD}\ pl}{\dfrac{\Delta \Rightarrow pl : \exists \mathrm{LexD.A}}{\Delta \Rightarrow pl : \mathrm{BR}}\,inc-R}\,\exists - R \quad \dfrac{\dfrac{\exists \mathrm{LexD.A} \Rightarrow \exists \mathrm{LexD.A}}{\exists \mathrm{LexD.A} \Rightarrow \mathrm{PIL_{BR}}}\sqcup\text{-R} \quad \dfrac{\Omega}{\mathrm{PIL_{BR}} \Rightarrow \mathrm{BR}}}{\exists \mathrm{LexD.A} \Rightarrow \mathrm{BR}}\,cut$$

$$\dfrac{\Delta \Rightarrow ml : \mathrm{BR} \quad \dfrac{\dfrac{\Pi}{\Delta \Rightarrow pl : \mathrm{BR}} \quad \dfrac{\Omega}{\boxed{ml : \mathrm{BR}, pl : \mathrm{BR} \Rightarrow cmp : \mathrm{BR}}}}{\Delta, ml : \mathrm{BR} \Rightarrow cmp : \mathrm{BR}}\,cut}{\Delta \Rightarrow cmp : \mathrm{BR}}\,cut$$

O que é PLN? [1]

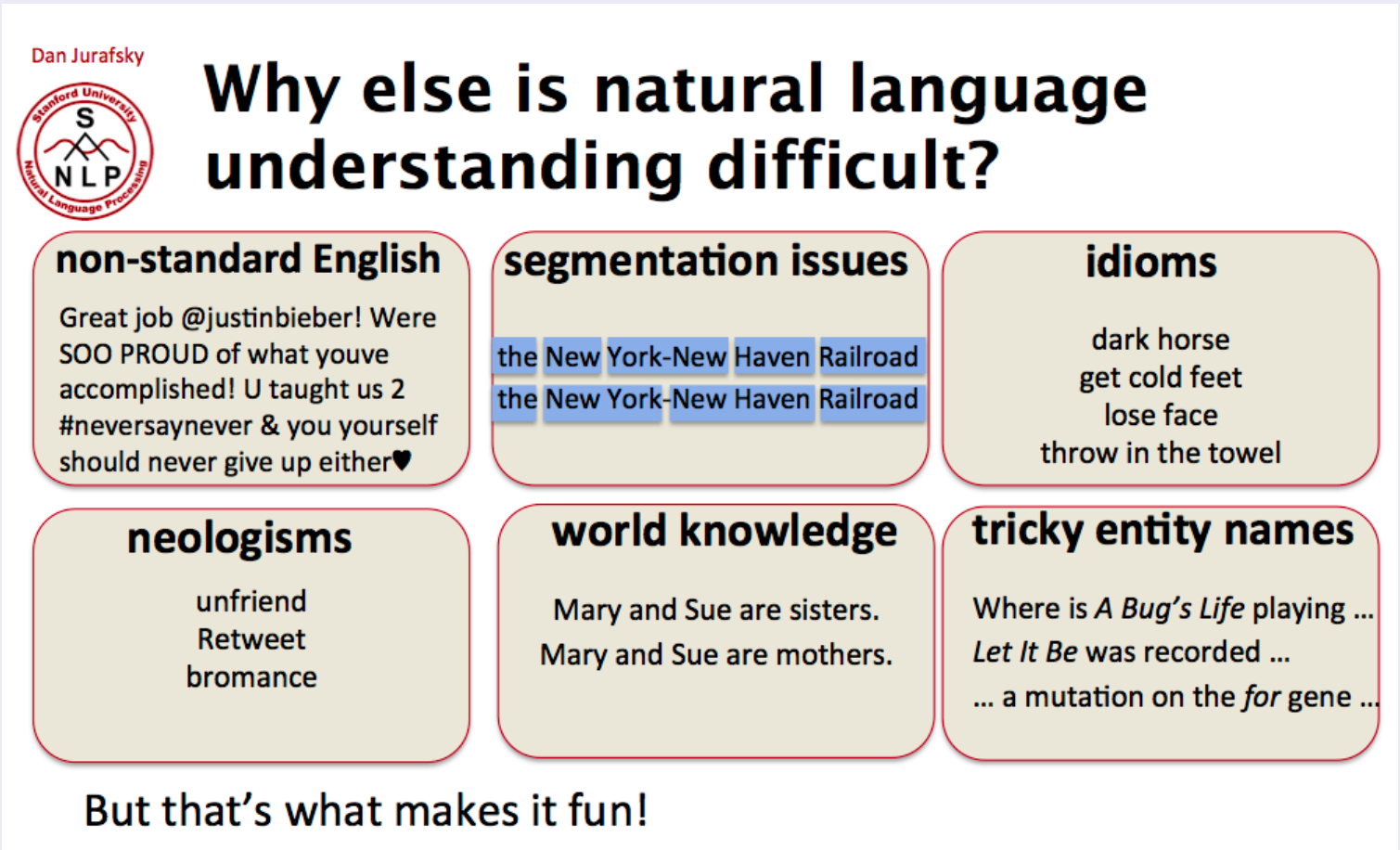
- Resposta à perguntas (IBM Watson ganhou o Jeopardy 2011)
- Extração de Informações (eventos e telefones de emails)
- Expanção de queries (via sinômimos)
- Análise de sentimentos (críticas em blogs e em sites online)
- Tradução
- Classificação ou agrupamento de textos
- Sumarização
- Linguagens controladas . . .

Ambiguidade é difícil!

- Em inglês: “Red Tape Holds Up New Bridges”.
- Em português: “João viu a bela mulher na rua com o binóculo.”.

[1] File intro-nlp.pdf em `https://class.coursera.org/nlp/`.

NLP é difícil



NLP é difícil

O que precisamos?

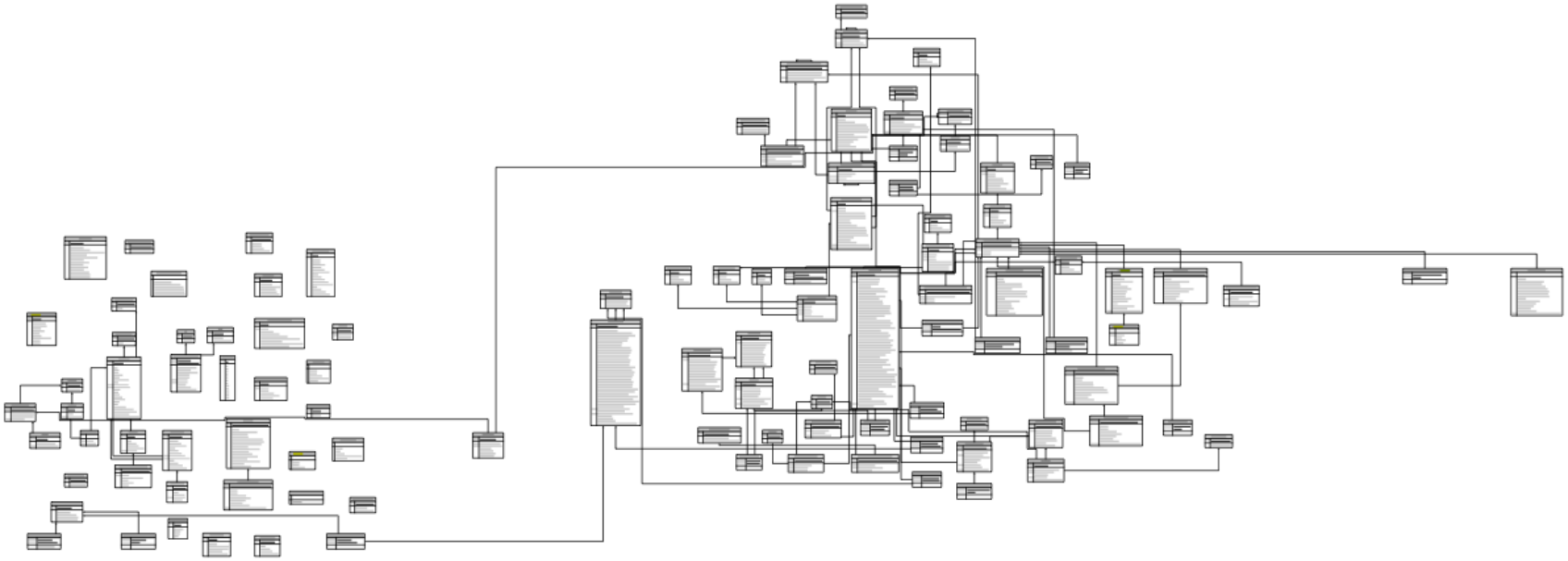
- Precisamos de informações sobre o mundo.
- Precisamos de informações sobre o idioma.
- Combinar conhecimento sobre idioma e mundo!

O projeto: PLN dos textos da histórica comtemporânea do Brasil

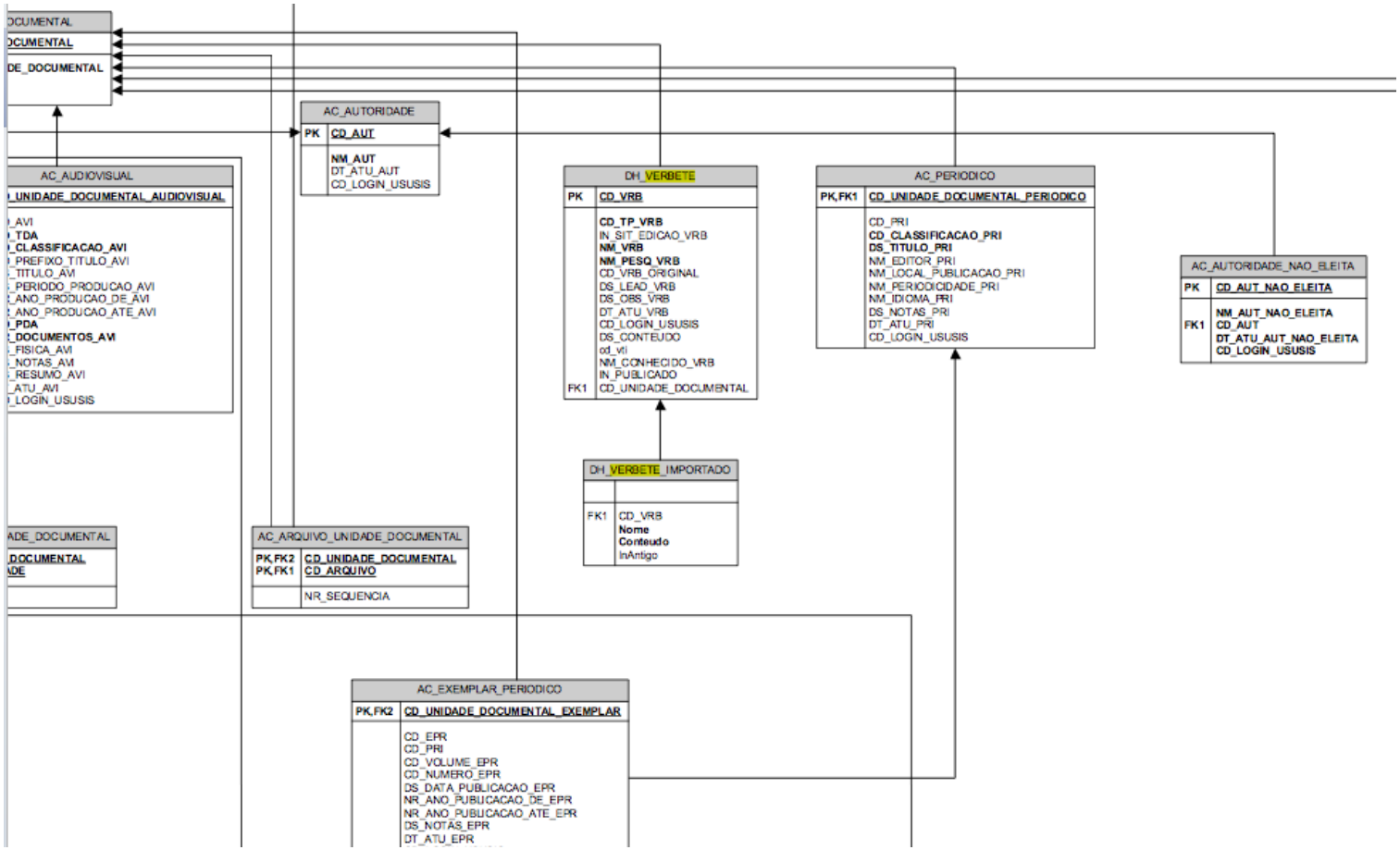
- No longo prazo, ferramentas lógicas para extração de conhecimento dos textos.
- Melhorar a estrutura das informações e capacidade de responder perguntas do sistema. Inferência de relações e propriedades implícitas sobre conceitos e termos.
- No contexto do MIST, foco no DHBB.

Preparando o terreno para usar os dados...

CPDOC: modelo ER de dados



CPDOC: modelo ER de dados



BD relacionais

- Rigidez para mudanças frequentes (diárias, semanais). Definições à priori.
- Tabelas adicionais para “guardar” relações muitos-muitos.
- Performance depende de decisões e manutenção de um DBA.
- Poucas restrições sobre o domínio no modelo.

+ Ferramentas disponíveis para desenvolvimento de sistemas de Informação. Padrões.
+ Disponibilidade de mão-de-obra.

“Selecting the next database for your project”, `http://www.franz.com`.

graph BD (triplestores)

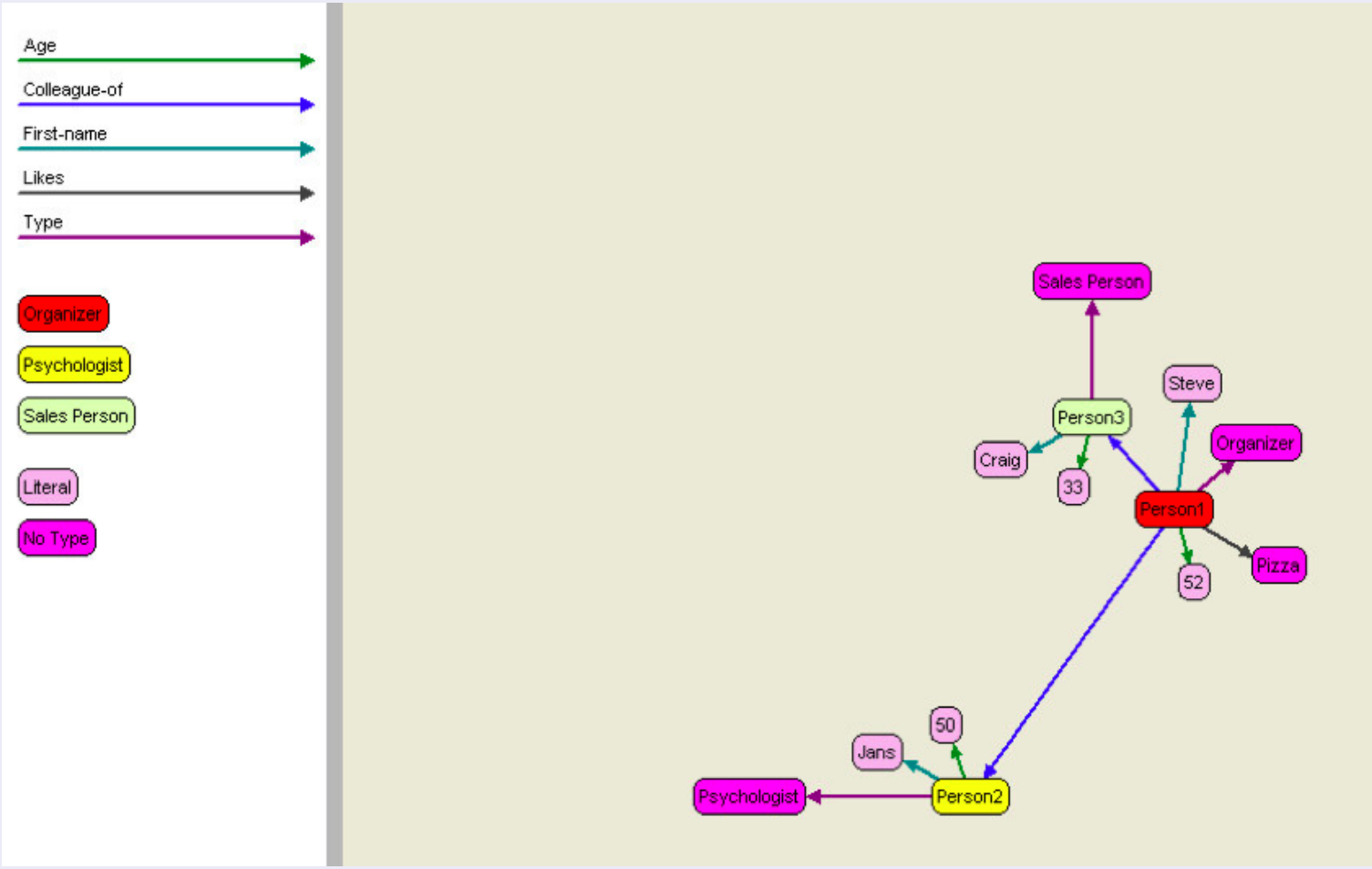
Triplas

```
createTripleStore(seminar.db)

addTriple (Person1 first-name Steve)
addTriple (Person1 isa Organizer)
addTriple (Person1 age 52)
addTriple (Person2 first-name Jans)
addTriple (Person2 isa Psychologist)
addTriple (Person2 age 50)
addTriple (Person3 first-name Craig)
addTriple (Person3 isa SalesPerson)
addTriple (Person3 age 32)

addTriple (Person1 colleague-of Person2)
addTriple (Person1 colleague-of Person3)

addTriple (Person1 likes Pizza)
```

"Selecting the next database for your project", `http://www.franz.com`.

graph BD (triplestores)

Grafo



graph BD (triplestores)

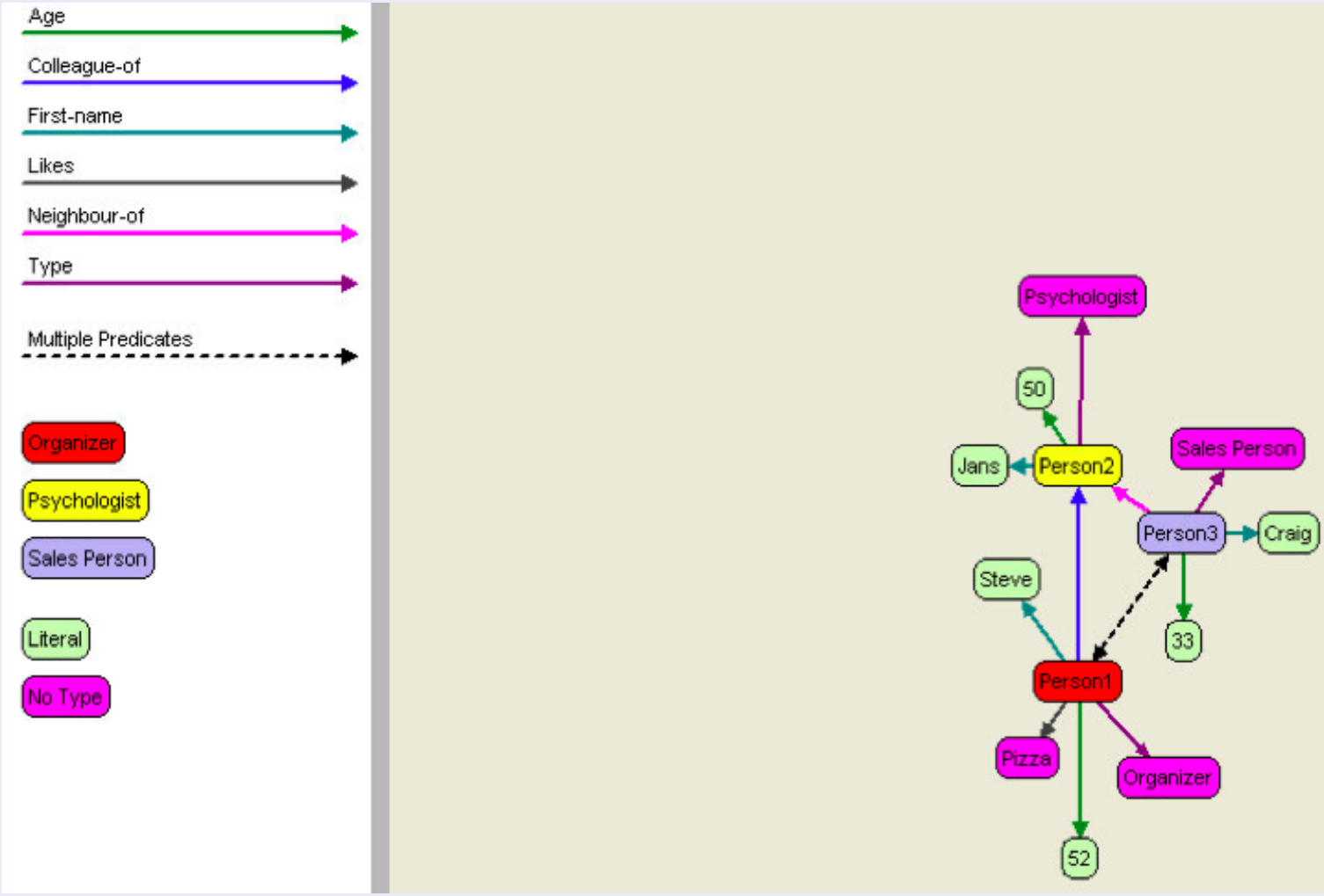
Novos “fatos”

```
addTriple ( Person3 neighbour-of Person1)
addTriple ( Person3 neighbour-of Person2)
```

“Selecting the next database for your project”, `http://www.franz.com`.

graph BD (triplestores)

Novo modelo



Graph databases

Vantagens

- Modelagem de diferentes tipos com diferentes propriedades.
- Expansível.
- Requisitos do domínio implementados por regras ou axiomas, no modelo.
- Queries complexas
- Protocolos e Padrões: SPARQL, OWL, RDF, RDFS etc.
- Fácil interoperabilidade.

Graph databases

Consultas

Find all meetings that happened in November within 5 miles of Berkeley that was attended by the most important person in Jans' friends and friends of friends.

```
(select (?x)
   (ego-group person:jans knows ?group 2)                    SNA
   (actor-centrality-members ?group knows ?x ?num)           SNA
   (q ?event fr:actor ?x)                                    DB Lookup
   (qs ?event rdf:type fr:Meeting)                           RDFS
   (interval-during ?event "2008-11-01" "2008-11-06")        Temporal
   (geo-box-around geoname:Berkeley ?event 5 miles)          Spatial
   !)
```

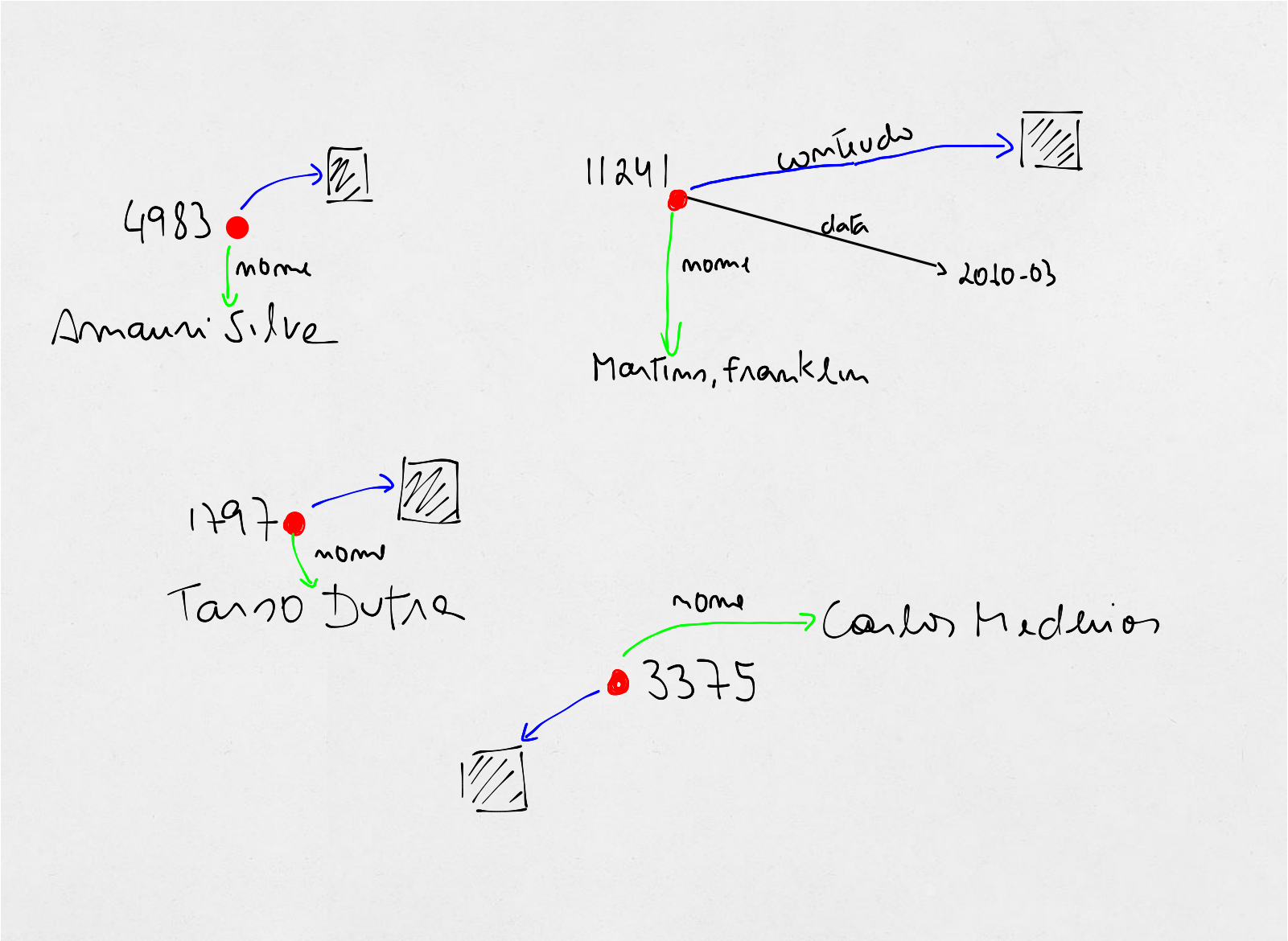
"Selecting the next database for your project", `http://www.franz.com`.

Graph databases

ER → Graph

- Fácil! Ferramenta `http://d2rq.org/d2r-server`!
- Mas... Ajustes no modelo são necessários!
- Vide exemplo

DBHH como um grafo



DBHH como um grafo

Vantagens do modelo Grafo

- Fácil integração de vocabulários e modelos.
- Fácil armazenamento de resultados (novas propriedades)
- Interoperabilidade entre sistemas.

LSA: primeiro exercício

- LSA tutorial.
- Limitações do LFA. Wikipedia
- Precisava do DHBB em arquivos...

DBHH como um grafo

Protótipo

- Mostrar protótipo
- Mostrar arquivos
- Idéias: (1) 1 verbete $\rightarrow$ 1 arquivo (URL e RDF); (2) Solr; (3) Geração de site Estático; (4) DVC (git system).
- Desvantagem: feedbacks não são incorporados ao DHBB.

Voltando ao problema principal...

NLP: o que precisamos?

Passos básicos não tão triviais:

- Importação de documentos (HTML, PDF etc)
- Tokenização (ex: “Dr. Fulano da F.G.V.”)
- Remoção de palavras não desejadas (stop words)
- Stemming (ex: educado, educada, educados etc. $\rightarrow$ educad)
- Lemmatization (ex: educar. A entrada do dicionário.)

Thanks Gerard de Melo.

NLP: o que precisamos também...

Queremos aproveitar ferramentas para o inglês. Mas precisamos de informações sobre o (em) português.

- Wordnet-like dicionário.
- Named entity reconizer.
- SUMO para o Português.
- NOMLEX-BR
- Verbnet-like KB.
- FrameNet-like KB.
- Gramática para o português (LFGs for XLE?)

Agenda de pesquisa inicial

- Construir uma Wordnet em português.
- Conectar a Wordnet-PT à SUMO Ontology para: (1) usar a SUMO; (2) conexão da Wordnet-PT com outras Wordnets.
- Investigar o uso da Wordnet-PT para resolução de ambiguidades: (1) expansão de consultas; (2) subjunção de textos.
- No DHBB: (1) extração de entidades nomeadas; e (2) extração de relações entre entidades (parentesco, amizade etc.)

Wordnet: o que e para que?

O que?

- Uma espécie de dicionário.
- Palavras são agrupadas em synsets (conjuntos = conceitos). Sinôminos em um dado contexto.
- Synsets são relacionados (rel. semânticas) e palavras são relacionadas (rel. sintáticas).

Wordnet: o que e para que?

Synset

WORDS	mouth, speak, talk, utter, verbalise, verbalize
GLOSS	express in speech
EXAMPLE	"She talks a lot of nonsense";
EXAMPLE	"This depressed patient does not verbalize"

Princeton WordNet online

Wordnet: o que e para que?

Para que?

- Word Sense Disambiguation (expresso pode ser: (1) explícito; (2) rápido; ou (3) verbo expressar).
- Expansão de consultas

A OpenWordnet-PT

- Disponível para download
- Open Multilingual Wordnet. Vide estatísticas. Exemplo de consulta.
- 7422 adjetivos, 55951 nomes, 1726 advérbios e 7155 verbos.
- Cobertura? Comparando com o DHBB? (1) Lemmatization; (2) Remoção de stop words.
- Correção? Verificação manual vide templates de frases.

Vide arquivos. Exemplos do DHBB:

- transmite, transmitiam, transmitira, transmitirem $\rightarrow$ transmitir.
- tolerado, tolerando, toleraria, tolerariam, toleráveis $\rightarrow$ tolerar.
- Estado (2979 vezes) $\rightarrow$ estar?
- Ingressou (182), reingressou (7)e ingressou (745) $\rightarrow$ ingressar

A OpenWordnet-PT

- Disponível para download
- Open Multilingual Wordnet. Vide estatísticas. Exemplo de consulta.
- 7422 adjetivos, 55951 nomes, 1726 advérbios e 7155 verbos.
- Cobertura? Comparando com o DHBB? (1) Lemmatization; (2) Remoção de stop words.
- Correção? Verificação manual vide templates de frases.

Vide arquivos. Exemplos do DHBB:

- transmite, transmitiam, transmitira, transmitirem $\rightarrow$ transmitir.
- tolerado, tolerando, toleraria, tolerariam, toleráveis $\rightarrow$ tolerar.
- Estado (2979 vezes) $\rightarrow$ estar?
- Ingressou (182), reingressou (7)e ingressou (745) $\rightarrow$ ingressar

A OpenWordnet-PT

- Disponível para download
- Open Multilingual Wordnet. Vide estatísticas. Exemplo de consulta.
- 7422 adjetivos, 55951 nomes, 1726 advérbios e 7155 verbos.
- Cobertura? Comparando com o DHBB? (1) Lemmatization; (2) Remoção de stop words.
- Correção? Verificação manual vide templates de frases.

Vide arquivos. Exemplos do DHBB:

- transmite, transmitiam, transmitira, transmitirem $\rightarrow$ transmitir.
- tolerado, tolerando, toleraria, tolerariam, toleráveis $\rightarrow$ tolerar.
- Estado (2979 vezes) $\rightarrow$ estar?
- Ingressou (182), reingressou (7)e ingressou (745) $\rightarrow$ ingressar

A OpenWordnet-PT

- Disponível para download
- Open Multilingual Wordnet. Vide estatísticas. Exemplo de consulta.
- 7422 adjetivos, 55951 nomes, 1726 advérbios e 7155 verbos.
- Cobertura? Comparando com o DHBB? (1) Lemmatization; (2) Remoção de stop words.
- Correção? Verificação manual vide templates de frases.

Vide arquivos. Exemplos do DHBB:

- transmite, transmitiam, transmitira, transmitirem $\rightarrow$ transmitir.
- tolerado, tolerando, toleraria, tolerariam, toleráveis $\rightarrow$ tolerar.
- Estado (2979 vezes) $\rightarrow$ estar?
- Ingressou (182), reingressou (7)e ingressou (745) $\rightarrow$ ingressar

A OpenWordnet-PT

- Disponível para download
- Open Multilingual Wordnet. Vide estatísticas. Exemplo de consulta.
- 7422 adjetivos, 55951 nomes, 1726 advérbios e 7155 verbos.
- Cobertura? Comparando com o DHBB? (1) Lemmatization; (2) Remoção de stop words.
- Correção? Verificação manual vide templates de frases.

Vide arquivos. Exemplos do DHBB:

- transmite, transmitiam, transmitira, transmitirem $\rightarrow$ transmitir.
- tolerado, tolerando, toleraria, tolerariam, toleráveis $\rightarrow$ tolerar.
- Estado (2979 vezes) $\rightarrow$ estar?
- Ingressou (182), reingressou (7)e ingressou (745) $\rightarrow$ ingressar

Correção da OpenWordnet-PT

Idéias

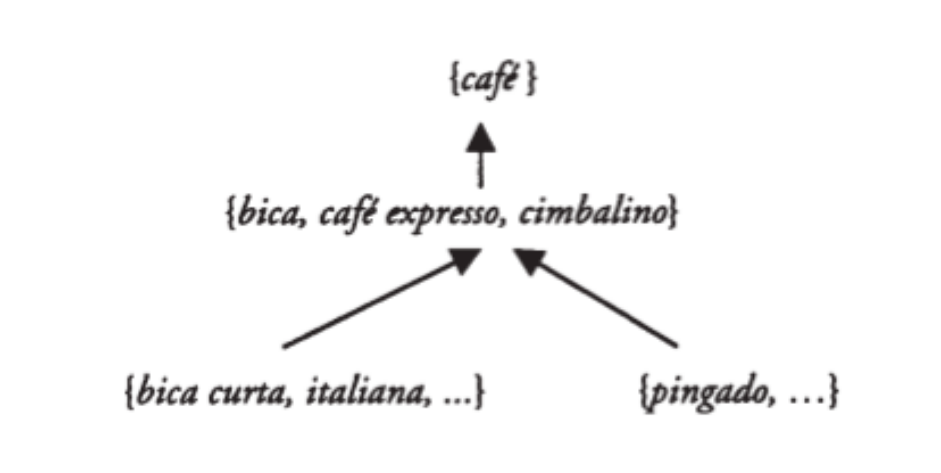
- Via template de sentenças? (EuroWordNet project).
- Se A e B são sinônimos, simetria é requerida. Teste 1: “A é B | B é A”. Teste 2: “A é um tipo de B | B é um tipo de A”.
- Se A é hipônimo de B. Teste: “A é um tipo de B” e “B não é um tipo de A”.
- Outros testes.

Correção da OpenWordnet-PT

Exemplos

- “Uma bica é uma bebida.” (Verdade)
- “Uma bebida é uma bica.” (Falso)
- “Uma bica é um expresso.” (Verdade)

Portuguese Wordnet: General archicteture and Internal Semantic Relations by Palmira Marrafa.



SUMO Ontology

- SUMO é uma ontologia de topo (conjunto de definições em uma Ling formal).
- Uma tentativa de capturar os mais gerais e reusáveis termos e definições.
- Termos da SUMO foram mapeadas para a synsets da WordNet.
- Algumas palavras são “vagas” para uma definição formal.
- Sigma Interface

Thanks Adam Pease

SUMO Ontology

- SUMO é uma ontologia de topo (conjunto de definições em uma Ling formal).
- Uma tentativa de capturar os mais gerais e reusáveis termos e definições.
- Termos da SUMO foram mapeadas para a synsets da WordNet.
- Algumas palavras são “vagas” para uma definição formal.
- Sigma Interface

Thanks Adam Pease

SUMO Ontology

- SUMO é uma ontologia de topo (conjunto de definições em uma Ling formal).
- Uma tentativa de capturar os mais gerais e reusáveis termos e definições.
- Termos da SUMO foram mapeadas para a synsets da WordNet.
- Algumas palavras são “vagas” para uma definição formal.
- Sigma Interface

Thanks Adam Pease

SUMO Ontology

- SUMO é uma ontologia de topo (conjunto de definições em uma Ling formal).
- Uma tentativa de capturar os mais gerais e reusáveis termos e definições.
- Termos da SUMO foram mapeadas para a synsets da WordNet.
- Algumas palavras são “vagas” para uma definição formal.
- Sigma Interface

Thanks Adam Pease

SUMO Ontology

- SUMO é uma ontologia de topo (conjunto de definições em uma Ling formal).
- Uma tentativa de capturar os mais gerais e reusáveis termos e definições.
- Termos da SUMO foram mapeadas para a synsets da WordNet.
- Algumas palavras são “vagas” para uma definição formal.
- Sigma Interface

Thanks Adam Pease

SUMO vs. WordNet

- “bright” como “full of promise”.
- “John has a bright future. He was selected for the varsity basketball team as a freshman.”
- Em outro contexto, “John is bright”... Ele provavelmente será eleito presidente...
- A palavra “walk”? Mais fácil ter definição formal e ser organizada em uma hierarquia de movimentos.

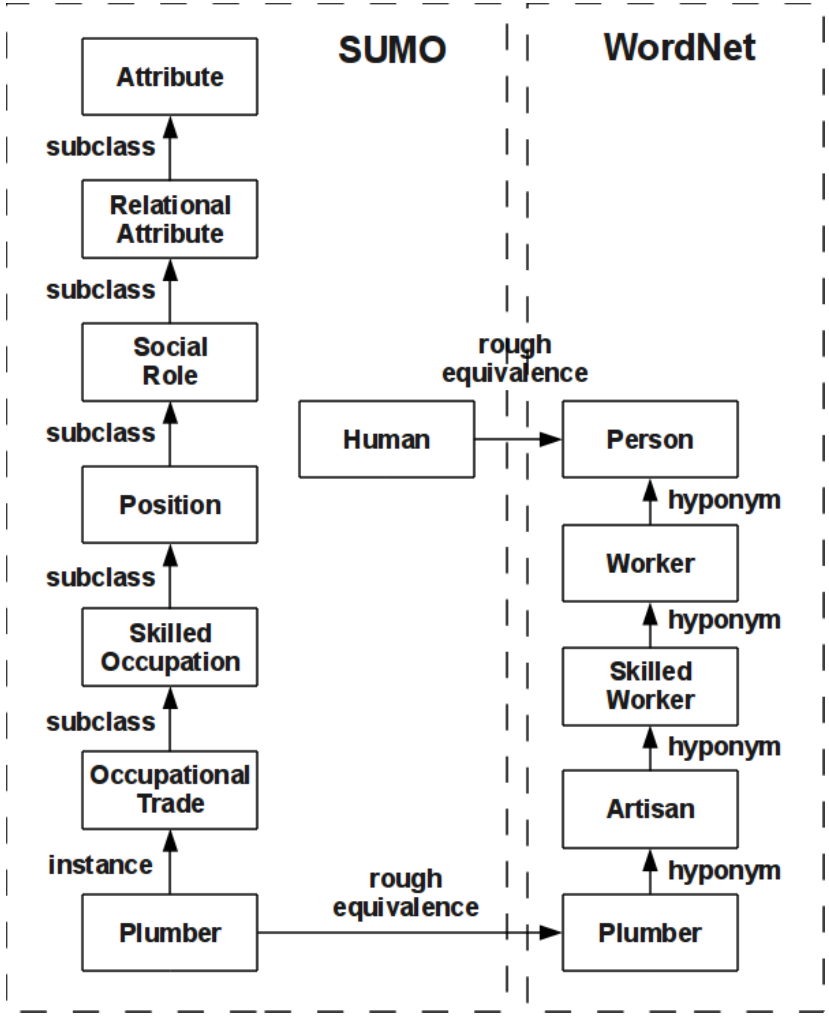
Thanks Adam Pease

SUMO vs. WordNet

- SUMO é uma ontologia: (1) regras; (2) formal; (3) para ser usado por um provador de teoremas. Feita para ser consistente.
- Wordnet é uma base de dados léxica.
- Léxicos são coletados e não podem ser livremente criados.
- Palavras podem ser vagas e ambíguas (Para que?).
- “transient role” vs. tipo.
- Wordnet é usada para modelar uma linguagem
- SUMO é usada para modelar a realidade.
- A conexão de ambos os recursos permite melhor entender a linguagem.

Thanks Adam Pease

SUMO vs. Wordnet



Thanks Adam Pease

SUMO e Português

- Extender SUMO com definições da cultura brasileira.
- Mapeamento da SUMO para a OpenWordNet-PT: conceitos não lexicalizáveis em inglês.
- Exemplo: churrascaria?!

Definição formal de Churrascaria?

```
(subclass MeatRestaurant Restaurant)
(=>
 (and
  (instance ?X MeatRestaurant)
  (instance ?F Meal)
  (located ?F ?X))
 (and
  (equals ?P1
    (ProbabilityFn
      (exists (?FM)
        (and (instance ?FM Meat)
             (contains ?F ?FM))))
  (equals ?P2
    (ProbabilityFn
      (not
        (exists (?FM)
          (and (instance ?FM Meat)
               (contains ?F ?FM))))
  (greaterThan ?P1 ?P2)))
```

Thanks Adam Pease

Obrigado!

S: (v) thank, give thanks (express gratitude or show appreciation to)

```
(=>
  (and
      (instance ?THANK Thanking)
      (agent ?THANK ?AGENT)
      (patient ?THANK ?THING)
      (destination ?THANK ?PERSON))
  (and
      (instance ?PERSON Human)
      (or
          (holdsDuring
              (WhenFn ?THANK)
              (wants ?AGENT ?THING))
          (holdsDuring
              (WhenFn ?THANK)
              (desires ?AGENT ?THING)))))
```