

Extracção Automática de Ontologias a Partir de Texto

Parte I **Introdução**

O Plano a longo prazo....

- Apresentações (+/-) sistemáticas
 - Revisão Bibliográfica
 - Demonstrações de protótipos
 -
- Discussão de Ideias

Agenda

- Definições de Ontologias
- Tipos de Ontologias
- Extracção Automática de Ontologias a partir de Texto
 - Métodos ←
 - Avaliação
 - Aplicações

Ontologia – Definições

- **Filosóficas**

- Tenta identificar e categorizar tudo que existe.
 - O que caracteriza existir?
 - O que significa existir?
- Aristóteles – Primeiro sistema de classificação (**taxionomia**) que ordenou os animais pelo tipo de reprodução.

Ontologia – Definições

- Computacionais
 - Um artefacto constituído por um vocabulário específico para descrever uma certa realidade. E um conjunto de assunções sobre o significado de cada item do vocabulário.

Ontologia – Definições (Fensel)

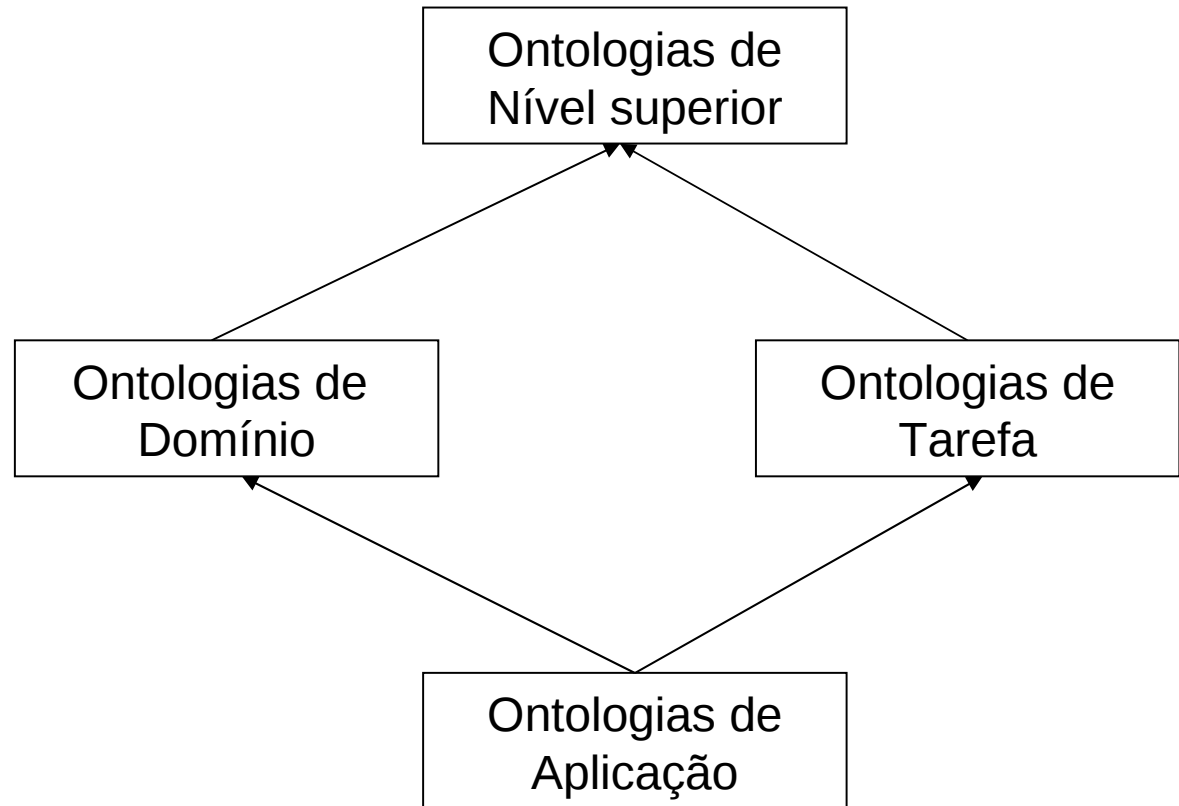
- Computacionais
 - É uma especificação **explícita** e **formal** de uma **conceptualização partilhada**.
 - Conceptualização – Os conceitos pertencentes ao domínio de interesse.
 - Explícito – O tipo de conceitos e as restrições de utilização estão explicitamente definidas.
 - Formal – “Machine Readable”.
 - Partilhada – Consensual e aceite por um grupo de pessoas.

Tipos de Ontologias (Buitelaar et al.)

Abstracto



Específico

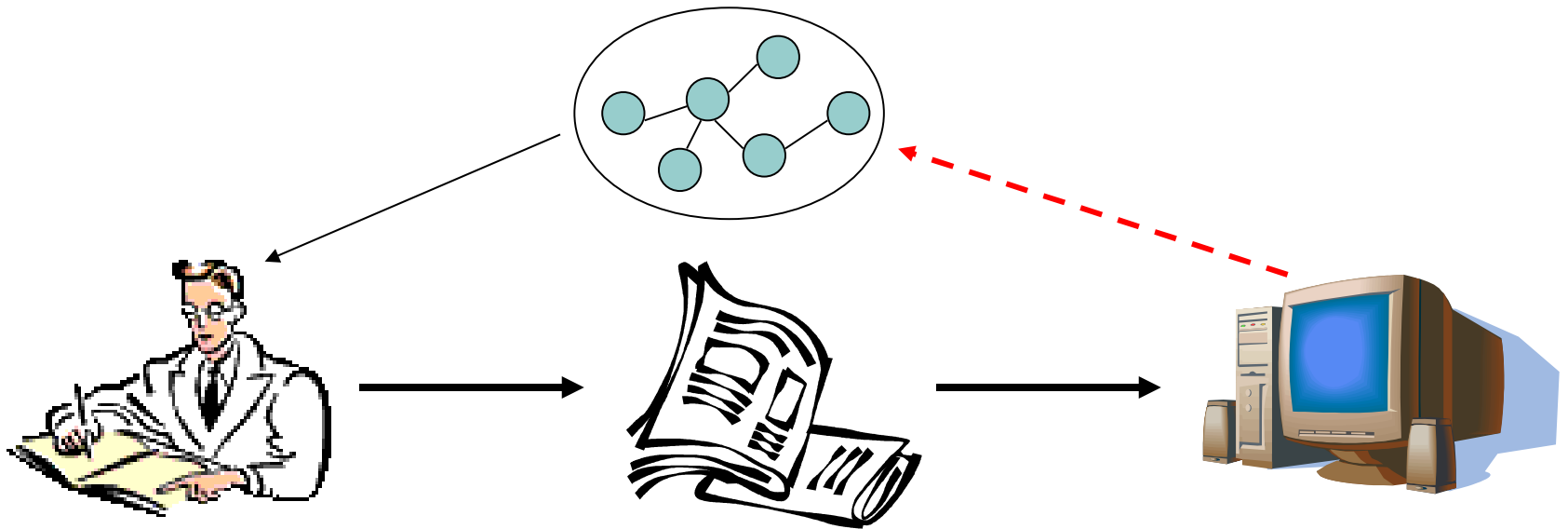


Extracção Automática de Ontologias a partir de Texto (EAOT)

Introdução

Extracção Automática de Ontologias a partir de Texto (EAOT)

- Pode ser encarado como um processo de “*reverse-engineering*”



EAOT- Pilha de Entidades (Buitelar et al.)

$\forall x, y : (\text{sofreDe}(x, y) \rightarrow \text{doente}(x))$

AXIOMAS

$\text{membro_de}(\text{médico}, \text{hospital})$

RELAÇÕES

$\text{é_uma}(\text{médico}, \text{pessoa})$

TAXONOMIA

$\text{doença} = \langle \text{Intensão}, \text{Extensão}, \text{Lemmas} \rangle$

CONCEITOS

$\{\text{doença}\}, \{\text{médico}, \text{doutor}\}$

SINONIMOS

$\text{doença}, \text{médico}, \text{doutor}$

TERMOS

Termos

Esta seção traz de volta um pouco da longa história do DCC. O DCC-Departamento de Cultura Científica do Centro Acadêmico Pereira Barretto (DCC/CAPB), órgão responsável pela representação e encaminhamento científico dos alunos da UNIFESP/EPM, fundado em 1937, atua junto aos alunos promovendo vários cursos extracurriculares, palestras, conferências e discussões de interesse à área médica.

Módulo Reconhecimento de Entidades Mencionadas

Sinónimos

- Podemos utilizar recursos lexicais; (e.g, WordNet)
- Abordagens estatísticas de co-ocorrência
 - Co-ocorrências de 2^a ordem (e.g, LSA)
 - (**carro**, [p1,p2,p3,p4])
 - (**automóvel**, [p1,p3,p4,p5])
 - **carro** e **automóvel** são sinónimos (??)

Conceitos

- **Doença**

- **Intenção:** “é um nome que se dá a todo um conjunto de sinais e sintomas que o corpo ou a pessoa apresenta.”
- **Extensão:**
 - **Cancro, Malária, Febre Amarela,...**
- **Lemmas:**
 - **Doença, ...**

Conceitos – Intensão (Navigli et al.)

- *festival* – “a day or period of time set aside for feasting and celebration”
- *jazz* – “a style of dance music popular in the 1920s; similar to New Orleans jazz but played by large bands”



- *jazz festival* – “**a kind of festival**, a day or period of time set aside for feasting and celebration, **related to jazz**, a style of dance music popular in the 1920s”

Conceitos - Extensão (Etzioni et al.)

- Procurar padrões léxico-sintáticos num corpus
 - ... doenças ***tais como***, [d1,d2,d3].....
 - ... actores ***tais como***, [a1,a2,a3].....

Conceitos - Lemmas

- Os métodos semelhantes aos utilizados para extracção de sinónimos.

Taxionomia

- Considerado a “coluna vertebral” de qualquer Ontologia.
 - Relações do tipo é_um (is_a).
- Exemplo de extracção:
 - Procura em corpora de padrões léxico-sintáticos indicando relações de “é_um” (Hearst et al.)

Relações

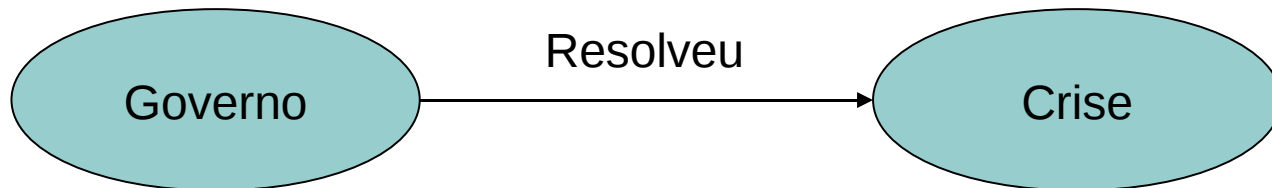
- Já temos:
 - Sinónimos
 - Hipónimos/Hiperónimos (“é_um”)
- Que outras relações modelar?
- Que nome dar à relação entre:
 - “**Companhia**” e “**Produto**”

Relações

- Podemos recorrer à utilização de padrões/heurísticas específicos:
 - KnowItAll
 - MindNet (Microsoft)
- Técnicas estatísticas (Kavavlec et al)
 - Procurar triplos(**Verbo_x**, **Conceito1**, **Conceito2**) numa janela de n palavras em texto.
 - Utilização de uma métrica, “*above expectation*”, para escolher o melhor verbo (etiqueta) para a relação.

Axiomas

- Servem para estabelecer equivalências entre relações. (bastante útil em *RAP*)



- Alguém encontrou solução para a crise?

Axiomas

- Procurar sintagmas que partilham o mesmo contexto. (Lin et al.)

<i>"X encontrou uma solução para Y"</i>		<i>"X resolveu Y"</i>	
comissão	greve	governo	problema
governo	crise	ela	mistério
ele	problema	investigador	problema
juiz	disputa	comissão	crise



encontrar uma solução para \approx resolver

Extracção Automática de Ontologias a Partir de Texto

FIM