Ontologias Linguísticas e Processamento de Linguagem Natural

Ygor Sousa – CIn/UFPE
ycns@cin.ufpe.br

2015

Roteiro

- Processamento de Linguagem Natural
- Ontologias Linguísticas
- WordNet
- FrameNet
- Desambiguação de Sentido

Processamento de Linguagem Natural

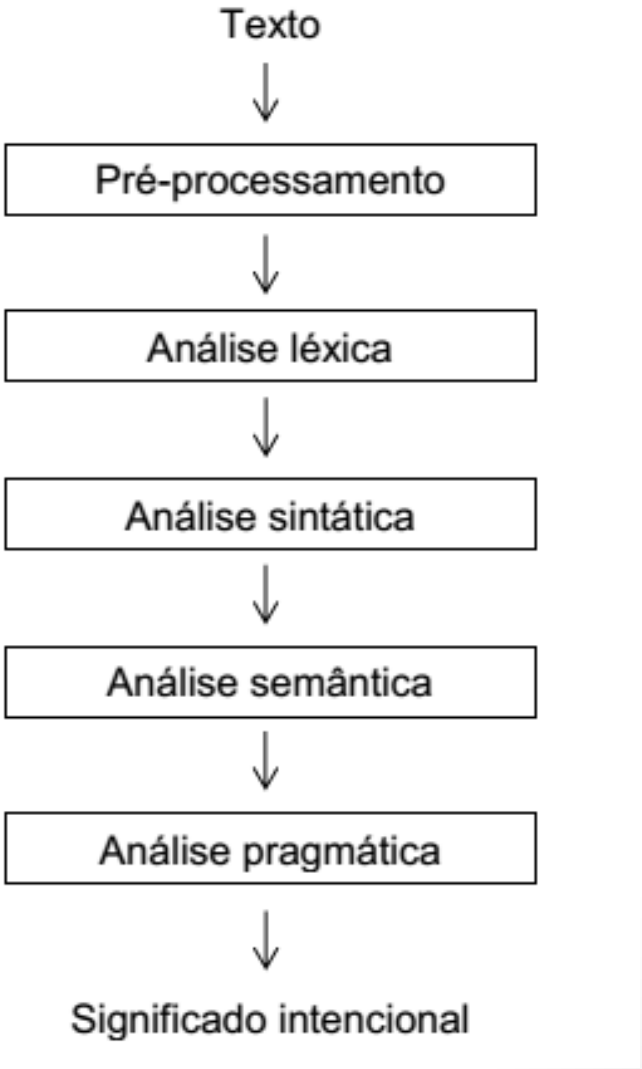
- São **técnicas** para analisar e representar naturalmente textos com o propósito de alcançar um **processamento de linguagem semelhante ao humano** em muitos diferentes tipos de atividades e aplicações [6].
- Textos ocorrem em um ou mais níveis de análise linguística.

Processamento de Linguagem Natural

- As técnicas de **PLN** são classificadas de acordo com o **nível de unidade** linguística processada [6]:
  - Nível Fonológico
  - Nível Morfológico
  - Nível Lexical
  - Nível Sintático
  - Nível Semântico
  - Nível de Discurso
  - Nível Pragmático
- **Raramente** um sistema de PLN aplica **todos os níveis**
- Maior o **nível**, maior a **complexidade** [6]

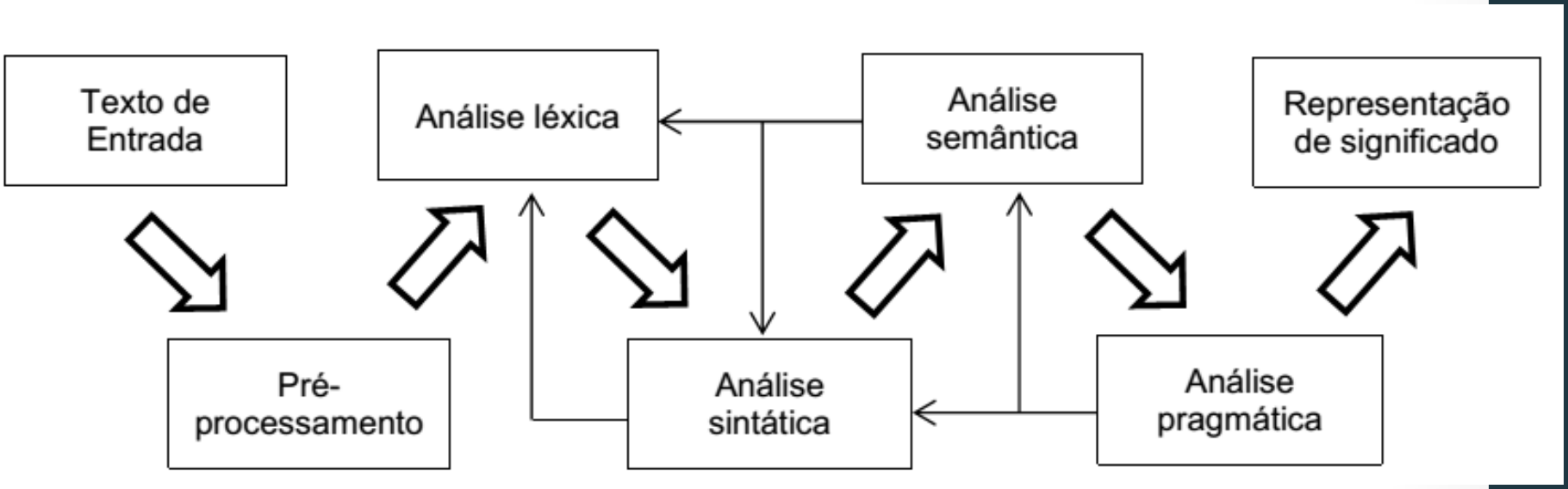
Processamento de Linguagem Natural

- Estágios de Análise [6]



Processamento de Linguagem Natural

- Modelo Retroalimentado [9]



Processamento de Linguagem Natural

- Diversos tipos de **aplicações** para **PLN** podem ser destacadas, dentre elas estão [8]:
  - Reconhecedores e Sintetizadores de Fala
  - Corretores Ortográficos e Gramaticais
  - Tradutores Automáticos
  - Geradores de Texto e Resumo
  - Extração de Informação
  - Interfaces de Linguagem Natural para Domínios Específicos

Ontologias Linguísticas

- Caracterizam-se por armazenar conceitos lexicalizados, isto é, conceitos expressos uma ou mais palavras de uma língua.
- Inventário de sentidos de conceitos compartilhados por uma comunidade linguística.
- Neste sentido, uma ontologia linguística em holandês, por exemplo, não armazenaria um conceito “container”, já que este não é lexicalizado nessa língua.

Ontologias Linguísticas

- Algumas das ontologias mais difundidas em PLN são:
  - **WordNet (http://wordnet.princeton.edu)**
  - SENSUS (http://www.isi.edu/natural-language/projects/ONTOLOGIES.html)
  - **FrameNet (https://framenet.icsi.berkeley.edu/)**
  - VerbNet (https://verbs.colorado.edu/~mpalmer/projects/verbnet.html)
  - Generalized Upper Model (http://www.ontospace.uni-bremen.de/ontology/gum.html)
  - Outros...

WordNet

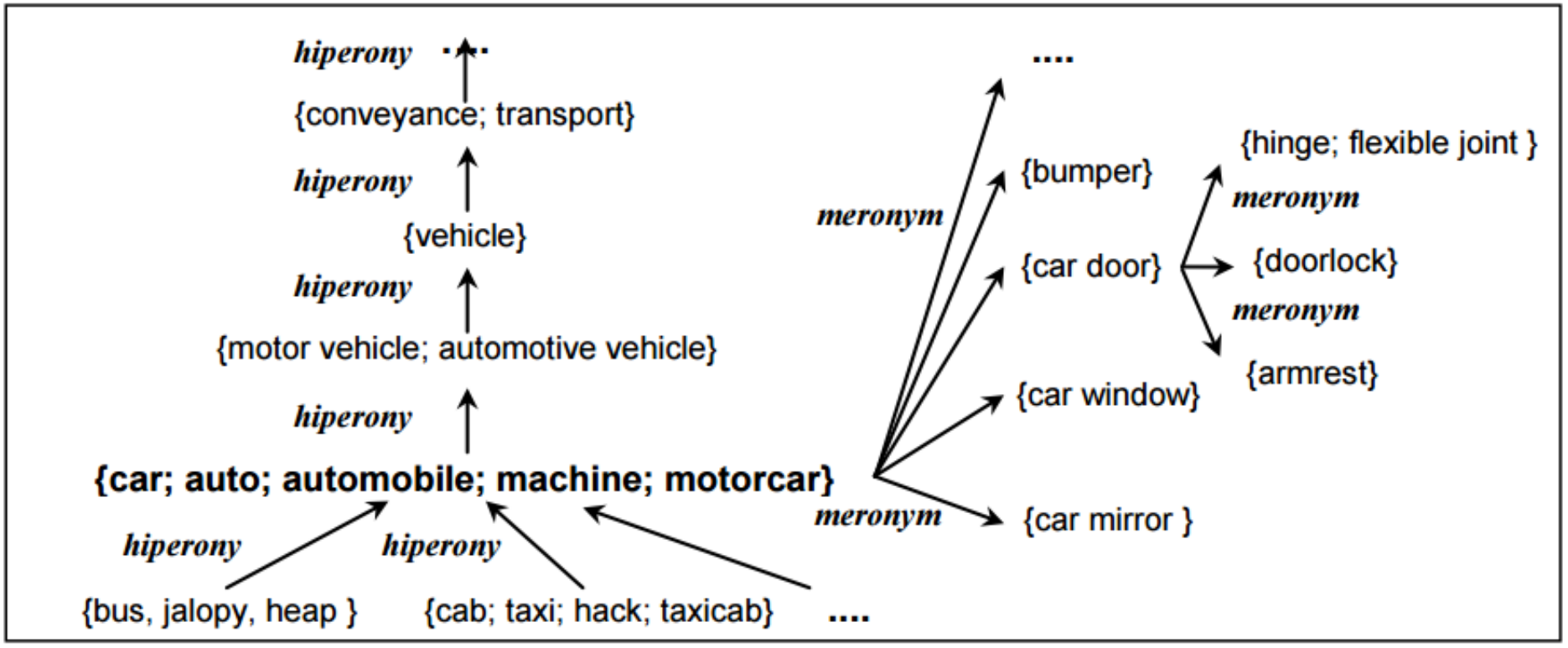
- Uma base de dados léxica organizada hierarquicamente
- Thesaurus + aspectos de um dicionário
  - Algumas outras línguas disponíveis ou em desenvolvimento
    - (Árabe, Finlandês, Alemão, Português…)
- Criado e mantido pela Universidade de Princeton

Categoria	Palavras Únicas
Noun	117,798
Verb	11,529
Adjective	22,479
Adverb	4,481

WordNet

- Organizada sob a forma de *Synsets* (conjunto de unidades sinônimas)
- Se relacionam por meio de relações lógico-conceituais como:
  - Hiperonímia / Hiponímia
  - Sinonímia / Antonímia
  - Holonímia / Meronímia

Exemplo abstrato de Synset de "car"



Sentido de “car” na Wordnet

Noun

- S: (n) **car**, auto, automobile, machine, motorcar (a motor vehicle with four wheels; usually propelled by an internal combustion engine) *"he needs a car to get to work"*
- S: (n) **car**, railcar, railway car, railroad car (a wheeled vehicle adapted to the rails of railroad) *"three cars had jumped the rails"*
- S: (n) **car**, gondola (the compartment that is suspended from an airship and that carries personnel and the cargo and the power plant)
- S: (n) **car**, elevator car (where passengers ride up and down) *"the car was on the top floor"*
- S: (n) cable car, **car** (a conveyance for passengers or freight on a cable railway) *"they took a cable car to the top of the mountain"*

WordNet

- O **synset** também instancia um conceito informal por ele lexicalizado, conhecido como gloss
- Exemplo: chump como um substantivo tem o gloss "a person who is gullible and easy to take advantage of"
- Esse sentido de "chump" é compartilhado por 9 palavras:
  - chump, fool, gull, mark, patsy, fall guy, sucker, soft touch, mug

Hierarquia de Hiperonímia da WordNet para "battery"

- S: (n) **battery**, electric battery (a device that produces electricity; may have several primary or secondary cells arranged in parallel or series)
  - *direct hyponym* / *full hyponym*
  - *part meronym*
  - *direct hypernym* / ***inherited hypernym*** / *sister term*
    - S: (n) electrical device (a device that produces or is powered by electricity)
      - S: (n) device (an instrumentality invented for a particular purpose) *"the device is small enough to wear on your wrist"; "a device intended to conserve water"*
        - S: (n) instrumentality, instrumentation (an artifact (or system of artifacts) that is instrumental in accomplishing some end)
          - S: (n) artifact, artefact (a man-made object taken as a whole)
            - S: (n) whole, unit (an assemblage of parts that is regarded as a single entity) *"how big is that part compared to the whole?"; "the team is a unit"*
              - S: (n) object, physical object (a tangible and visible entity; an entity that can cast a shadow) *"it was full of rackets, balls and other objects"*
                - S: (n) physical entity (an entity that has physical existence)
                  - S: (n) entity (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))

Relações de Substantivos da WordNet

Relation	Also called	Definition	Example
Hypernym	Superordinate	From concepts to superordinates	*breakfast*^{1} → *meal*^{1}
Hyponym	Subordinate	From concepts to subtypes	*meal*^{1} → *lunch*^{1}
Member Meronym	Has-Member	From groups to their members	*faculty*^{2} → *professor*^{1}
Has-Instance		From concepts to instances of the concept	*composer*^{1} → *Bach*^{1}
Instance		From instances to their concepts	*Austen*^{1} → *author*^{1}
Member Holonym	Member-Of	From members to their groups	*copilot*^{1} → *crew*^{1}
Part Meronym	Has-Part	From wholes to parts	*table*^{2} → *leg*^{3}
Part Holonym	Part-Of	From parts to wholes	*course*^{7} → *meal*^{1}
Antonym		Opposites	*leader*^{1} → *follower*^{1}

WordNet

- Onde está:
  - http://wordnet.princeton.edu/
  - http://wordnetweb.princeton.edu/perl/webwn (Online)
- Bibliotecas
  - Python: WordNet da NLTK
    - http://www.nltk.org/Home
  - Java:
    - JWNL, extJWNL no sourceforge

FrameNet

- Desenvolvido e mantido pelo International Computer Science Institute (ICST - Berkley).
- É uma base lexical para língua inglesa baseada na teoria de semântica de frames de que "significações são relativizadas a cenas" [2].
  - Mas já expandida para outras línguas como Alemão, Japonês, Francês, Espanhol e Português
- Assim, temos o frame como um "esquema imagético".

FrameNet

- Frames são compostos por Element Frames (EF) de diferentes classificações:
  - Nucleares;
  - Periféricos ou Não Nucleares;
  - Extratemáticos.
- Se conectam por relações como:
  - Herança
  - SubFrame
  - Causa de
  - Uso

Competition

Lexical Unit Index

Definition:

This frame is concerned with the idea that people (Participant_1, Participant_2, or Participants) participate in an organized, rule-governed activity (the Competition) in order to achieve some advantageous outcome (often the Prize). Rank and Score are different criteria by which the degree of achievement of the advantageous outcome is judged.

He and I PLAYED tennis.

FEs:

Core:

Competition [Comp]	This FE is used for the name of the competition. Jo PLAYED in the foosball tournament
Participant_1 [Partic-1] **Excludes:** Participants	This FE identifies the first (or only) participant in a competition. Jo PLAYED the lottery every day.
Participant_2 [Partic-2] **Requires:** Participant_1 **Excludes:** Participants	This FE identifies the second participant in a competition. Jo PLAYED Leslie at tennis.
Participants [Partic-S]	This FE is used for plural NP participants in a competition. Jo and Leslie PLAYED tennis.

Non-Core:

Degree []	This FE describes the intensity of competition.

The Mets will play at Shea Stadium tonight.

Frame-frame Relations:

Inherits from:
Is Inherited by:
Perspective on:
Is Perspectivized in:
Uses: Intentionally_act
Is Used by: Alliance, Sports_jargon
Subframe of:
Has Subframe(s): Finish_competition
Precedes:
Is Preceded by:
Is Inchoative of:
Is Causative of:
See also:

Lexical Units:

challenge.n, compete.v, competition.n, competitive.a, competitor.n, play.v, player.n, rival.n, rivalry.n

Created by MRLP on 11/21/2001 03:41:51 PST Wed

Lexical Unit	LU Status	Lexical Entry Report	Annotation Report	Annotator ID	Created Date
challenge.n	**Created**	Lexical entry	Annotation	JKR	07/18/2005 08:24:55 PDT Mon
compete.v	**Add_Annotation**	Lexical entry	Annotation	MJE	10/17/2003 10:30:58 PDT Fri
competition.n	**Finished_Initial**	Lexical entry	Annotation	JKR	11/21/2001 05:38:56 PST Wed
competitive.a	**Created**	Lexical entry	Annotation	JKR	11/21/2001 05:39:30 PST Wed

FrameNet

- Ao invés de palavras, o FrameNet trabalha com Lexical Units (LUs), cada uma sendo um par de palavra e sentido
  - Evitar polissemia, ambiguidade léxica, entre outros;
- Diferentes LUs em WordNet pertencem a diferentes synsets, em FrameNet (geralmente) pertencem a diferentes Frames

player.n

Frame: Competition

Definition:

FN: a person who is an active member of a competition

Frame Elements and Their Syntactic Realizations

The Frame Elements for this word sense are (with realizations):

Frame Element	Number Annotated	Realization(s)
Competition	(73)	A.Dep (1) DNI.-- (57) N.Dep (13) PP[on].Dep (1) Poss.Gen (1)
Participant_1	(21)	DEN.-- (21) NP.Appositive (1)
Participants	(63)	DEN.-- (63)
Place	(1)	PP[in].Dep (1)
Rank	(17)	A.Dep (1) AJP.Dep (16)
Venue	(1)	PP[on].Dep (1)

Valence Patterns:

These frame elements occur in the following syntactic patterns:

Number Annotated	Patterns			
13 TOTAL	Competition	Participant_1		
(10)	DNI	DEN		

FrameNet

- Frames são criados e legitimados a partir de textos selecionados em um corpus através de anotações.
  - Abaixo o Corpus do ano de 2010 no FrameNet Br;

GÊNERO	TOTAL
Oral	1.186.336
Didático	1.388.660
Jurídico	761.852
Literário	2.425.955
Téc. & Cien.	1.767.565
Jornalístico	27.203.360
Universitário	1.027.908
Mensagem eletrônica	1.202.297
Legendas de filmes	34.800.000
Total de Palavras: 71.763.933	

Annotation

Lexical Entry Competition

competitor.n

Frame Element	Core Type
Competition	Core
Degree	Peripheral
Duration	Peripheral
Frequency	Extra-Thematic
Manner	Peripheral
Means	Peripheral
Participants	Core
Participant_1	Core
Participant_2	Core
Place	Peripheral
Prize	Extra-Thematic
Purpose	Extra-Thematic
Rank	Peripheral
Score	Peripheral
Time	Peripheral
Venue	Peripheral

Turn Colors Off

- 250-s20-ppat
  1. Among the Polish runners is Antoni Niemczak , their fastest COMPETITOR at 2-9-41 , the fifth best in the race .
  2. The Mistral SST is the board used for the Youth and Ladies World Championships , for which identical boards are supplied for COMPETITORS at the race site .
  3. The carnival committee will be checking in the COMPETITORS at the village hall from 9.30 a.m .
- 250-s30-ppin
  1. WHEELCHAIR COMPETITORS in Sunday 's London Marathon have been told that instead of going for personal best times they must take at least one hour and 50 minutes to complete the course .
  2. But when the 28-year-old Luton osteopath *becomes* Britain 's only COMPETITOR in this month 's Pentjak Silat World

FrameNet

- Atualmente conta:
  - 1215 Frames;
  - 1827 Frame Relations;
  - 13308 Lexical Units (LUs);
  - 201226 Annotation sets.
- Onde encontrar:
  - https://framenet.icsi.berkeley.edu/fndrupal/home
  - https://framenet.icsi.berkeley.edu/fndrupal/index.php?q=frameIndex (Online)

WordNet/FrameNets: Aplicações

- Wordnets e FrameNets têm sido utilizadas nas seguintes tarefas/aplicações [4]:
  - Recuperação de informação;
  - Sumarização automática;
  - **Desambiguação de sentido;**
  - Categorização de textos,;
  - Tradução automática
  - Entre outras.

Desambiguação de Sentido

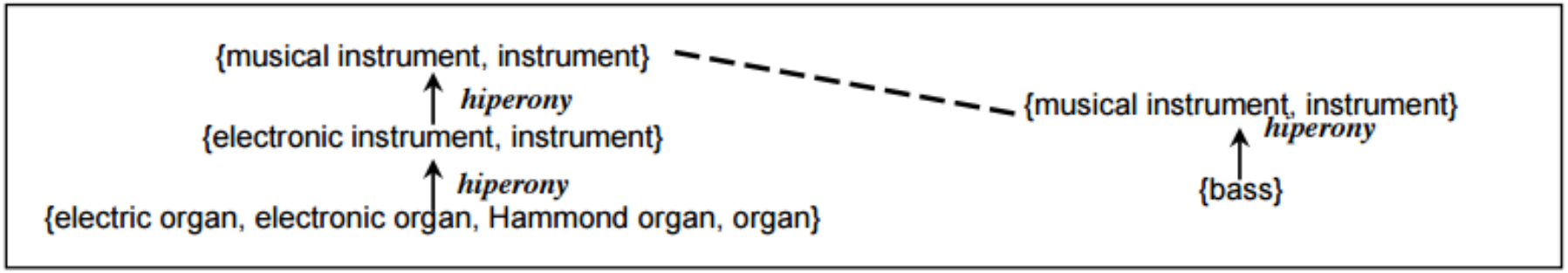
- Um dos problemas mais discutidos em PLN.
- Comum nas aplicações de PLN, como as outras discutidas anteriormente;
- Problema ocorre quando uma palavra apresenta mais de uma opção de sentido com a mesma categoria gramatical;
- Exemplo: “light”, pode ser “leve” ou “luz” (homonímia)
- Qual sentido escolher?

Desambiguação de Sentido

- De forma simples, a aplicação de WordNets nesta tarefa pode vir a partir da hipótese de que:
  - Palavras semanticamente relacionadas ou de um mesmo campo semântico tendem a co-ocorrer em um documento;
- Desta forma, a estratégia de modo geral pode ser:
  - Identificar os sentidos/synsets que contêm as palavras em foco;
  - Identificar as relações entre sentidos/synsets;
  - Identificar qual o sentido mais provável das palavras em foco;

Desambiguação de Sentido

- Exemplo simples: Uma frase tem as palavras “organ” e “bass”. O sistema identifica que elas estão em 6 e 8 synsets, respectivamente;



- Obtemos uma relação de hiperônimos entre synsets das palavras;
- Desta forma, podemos deduzir que o melhor sentido para as duas palavras é “instrumento musical”

Referências

- [1] Princeton University "About WordNet." WordNet. Princeton University. 2015. http://wordnet.princeton.edu
- [2] FILLMORE, C. Topics in lexical semantics. 1977.
- [3] International Computer Science Institute “FrameNet Documentation” FrameNet. Berkley. 2015 https://framenet.icsi.berkeley.edu/fndrupal/documentation
- [4] DI FELIPPO, Ariani. ONTOLOGIAS LINGÜÍSTICAS APLICADAS AO PROCESSAMENTO AUTOMÁTICO DAS LÍNGUAS NATURAIS: O CASO DAS REDES WORDNETS1. Múltiplas perspectivas em Linguística. Uberlândia: Edufu, 2008.
- [5] PAULO, H. FSI: Uma Infraestrutura de Apoio ao Projeto FrameNet Utilizando Web Semântica. 132f. Dissertação de Mestrado. Instituto de Ciências Exatas, Universidade Federal de Juiz de Fora, Recife, 2010.

Referências

- [6] LIDDY, E. D. Enhanced Text Retrieval Using Natural Language Processing. In: Bulletin of the American Society for Information Science, v. 24, n. 4, 1998.
- [7] VOSSEN, P. Ontologies. In: MITKOV, R. (Ed.). The Oxford handbook of Computational Linguistics. Oxford: Oxford University Press, 2003, p. 464-82
- [8] VIEIRA, R.; LIMA, V. L. S. Lingüística Computacional: Princípios e Aplicações. In: JAIA, SBC, Fortaleza, Brasil, 2001.
- [9] JUNIOR, A. T. FSI: Processamento de Linguagem Natural para Indexação Automática Semântico-Ontológica. 180f. Tese de Doutorado. Universidade de Brasília, Brasília, 2013.