

WordNet: Relações Semânticas e Métricas de Associação/Semelhança

Seminário Doutoral
Nuno Seco

Estrutura da Apresentação

- Relações de Semântica Lexical
 - Objecto de Estudo
 - WordNet
- Métricas de Semelhança no WordNet
 - Base de Conhecimento Lexical
 - Corpus
 - Teoria de Informação
 - Dicionários

WordNet

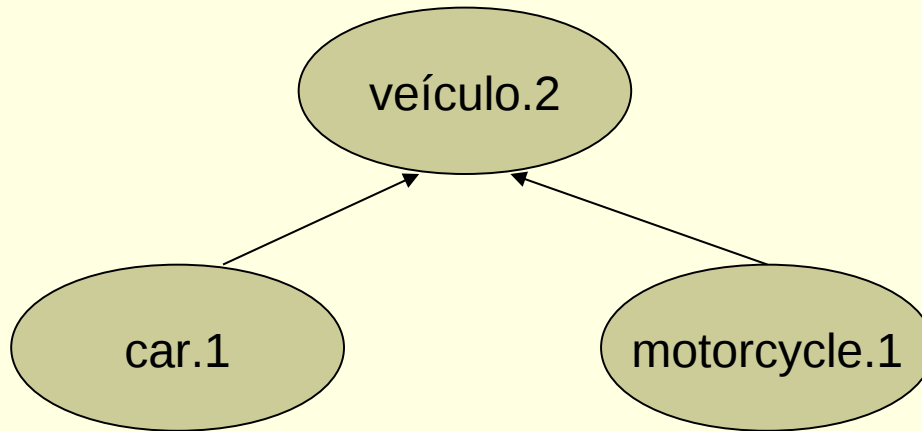
- É uma BCL inspirada em teorias psico-linguísticas.
 - Divisão em categorias sintáticas
 - Substantivos
 - Verbos
 - Advérbios
 - Adjectivos
 - Taxonomia de substantivos **estava(??)** particionada em 9 domínios diferentes. (evento, emoção, processo, etc)

Organização dos Termos

- Termos estão organizados em SynSets (Synonym Sets):
 - {car.1, auto.1, automobile.1, machine.1, motorcar.1}
 - a motor vehicle with four wheels; usually propelled by an internal combustion engine; "he needs a car to get to work"

Relações Semânticas

- As relações são estabelecidas entre synsets.



Relações Semânticas

- Hyperonímia/Hiponímia (substantivos, verbos)
- Meronímia (substantivos)
 - Substância
 - substância_de(lenhina, madeira)
 - Membro
 - membro_de(jogador, equipa)
 - Parte
 - parte_de(pata, gato)
- Sinonímia (todas as cat.)

Relações Semânticas

- Antonímia (todas as cat. “*lexical*”)
- Atributo (substantivo → adjetivo)
 - peso(leve), peso(pesado)
- Domínio (todas)
 - Categoria
 - topico_de(guerra, militar)
 - Região
 - região_de(saratoga, nova_york)

Relações Semânticas

- Causais (verbos)
 - causa(matar, morrer)
- Implicação (verbos)
 - Implica(ressonar, dormir)
- Derivação (adverbio → adjetivo, “*lexical*”)
 - derivado_de(somente, só)

Emprega uma visão de “Homônímia Forte”

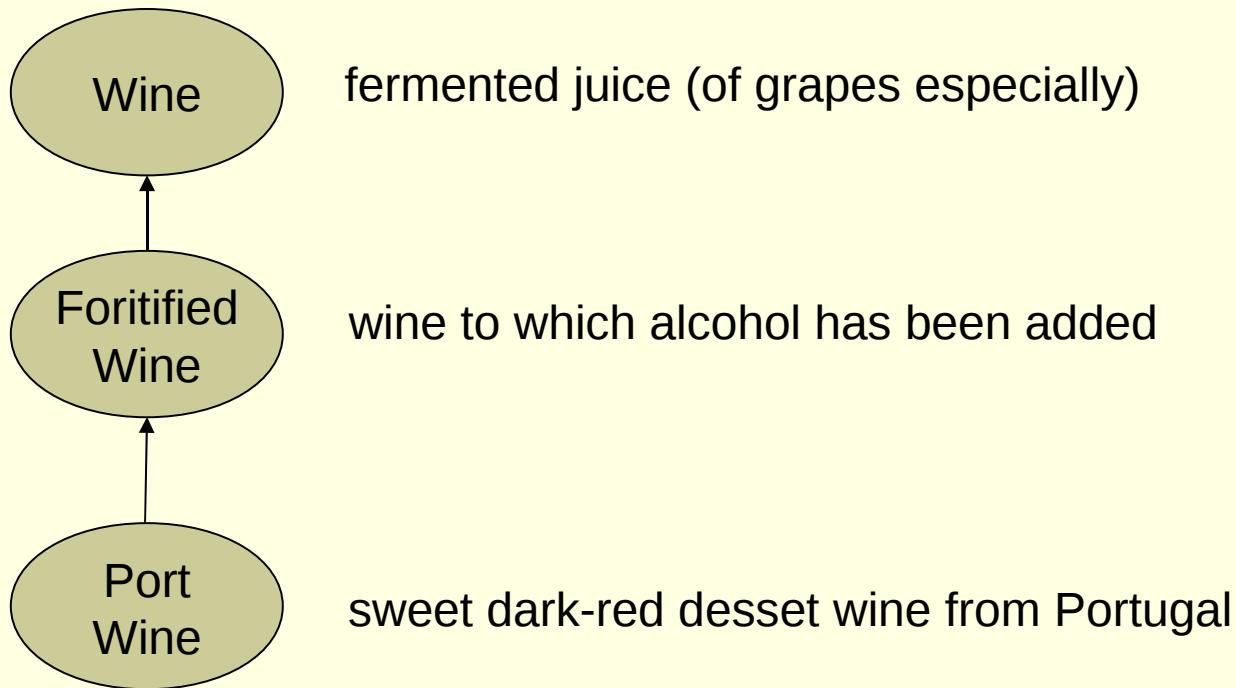
- {duck.1} -- small wild or domesticated web-footed broad-billed swimming bird usually having a depressed body and short legs.
- {duck.3} -- flesh of a duck (domestic or wild).
- Representam diferentes dimensões do mesmo conceito.

A Taxonomia

- A utilização de relações de hiperonímia é uma forte componente do WordNet.
 - 65% das relações (substantivos) são de hiperonímia/hiponímia
 - Permite uma estruturação eficiente dos conceitos.
 - Considere a organização de um super-mercado.

Teoria Diferenciadora

- A preocupação é fornecer atributos que distingam um conceito do seu hiperónimo.



Teoria Construtiva

- Um conjunto de conceitos primitivos.
 - São utilizados para construir novos conceitos
 - Exemplo:
 - HowNet -- Base de Conhecimento Lexical para o Chinês
 - YanJun
 - Yan –Sábio
 - Jun - Bonito
 - 800 conceitos primitivas → 110,000 conceitos

Associação Semântica

- Utilizado em motores de pesquisa como métrica de *“ranking”*.
- Utilização de algoritmos de Criatividade Computacional.
 - Geração de Conceitos
- Detecção de “Malapropisms”
 - Concerto vs. Conserto
 - Coro vs. Couro
 - Intercessão vs. Intersecção

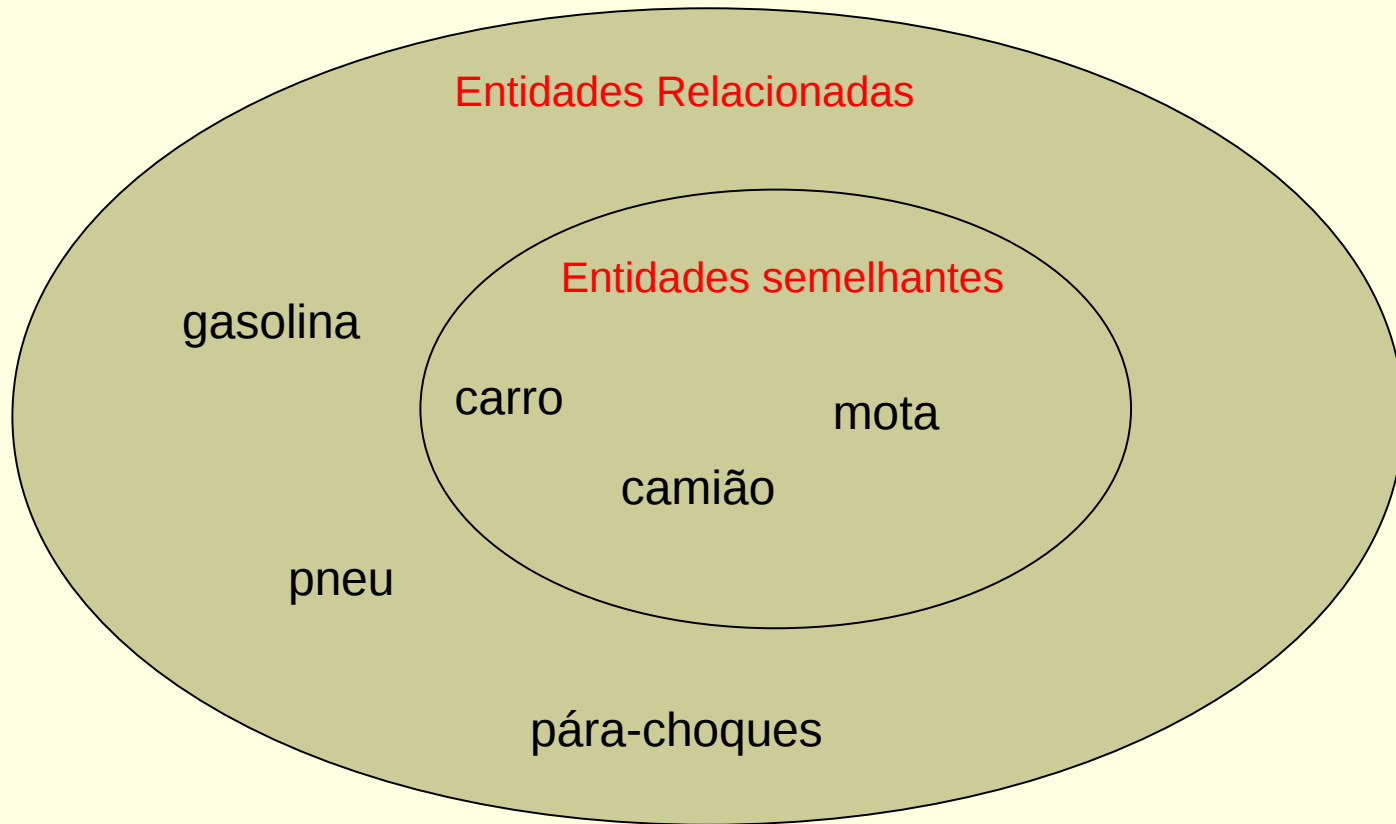
Semelhança e Associação Semântica

- São coisas diferentes mas normalmente não é feita a distinção na literatura.

Qual dos pares é mais semelhante?

- Carro --- Pára-choques
- Carro --- Bicicleta

Semelhança e Associação Semântica

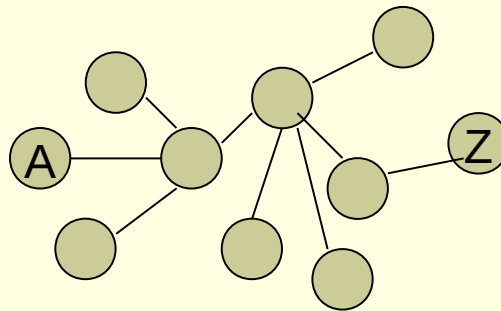


Tipos de Abordagens

- Baseado nas relações da BCL
- Baseado nas estatísticas derivadas de Corpus
- Baseado na “Teoria da Informação”
 - Abordagem híbrida (BCL, Corpus)
- Baseado em Dicionários (nas definições)

Base de Conhecimento Lexical

- A Base de Conhecimento pode encarada como um grafo.



- A associação semântica é calculado em função do número de arcos que separem dois conceitos.

Base de Conhecimento Lexical

- Alguns refinamentos a esta estratégia:
 - Só utilizar alguns tipos de relações
 - Por exemplo: Hyperonímia (semelhança)
 - Atribuição de pesos às relações

Baseado em Corpus

- Extracção de Co-ocorrências de palavras.
- Informação Mútua:

$$I(x, y) = \log_2 \left(\frac{P(x, y)}{P(x)P(y)} \right)$$

- Compara a probabilidade de x e y co-ocorrerem com a probabilidade de ocorrerem independentemente.

Baseado em Corpus

- Vector Space Model
 - Para cada palavra cria-se um vector contendo as frequências das palavras que co-ocorrem com a primeira.

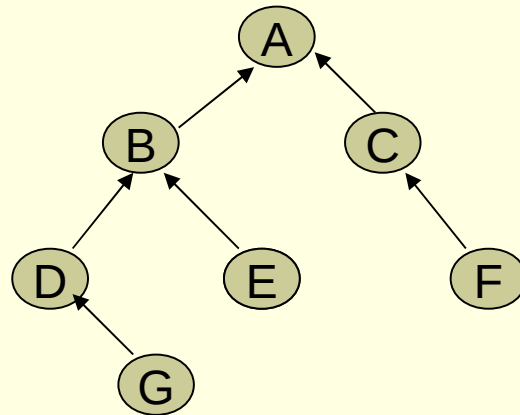
- Latent Semantic Analysis (LSA)

Teoria de Informação

- Na realidade são abordagens híbridas
 - Utilizam BCL
 - Corpus
 - Restringem-se às relações hiperonímia (semelhança).
 - Tentam quantificar a informação que um conceito expressa.
- Noção Base
 - Quantidade de Informação (*"Information Content"*)

$$IC(c) = -\log_2(P(c))$$

Teoria de Informação



$$P(A) = P'(A) + P(B) + P(C)$$

$$P(A) \approx 1 \rightarrow IC(A) \approx 0$$

$$P(B) = P'(B) + P(D) + P(E)$$

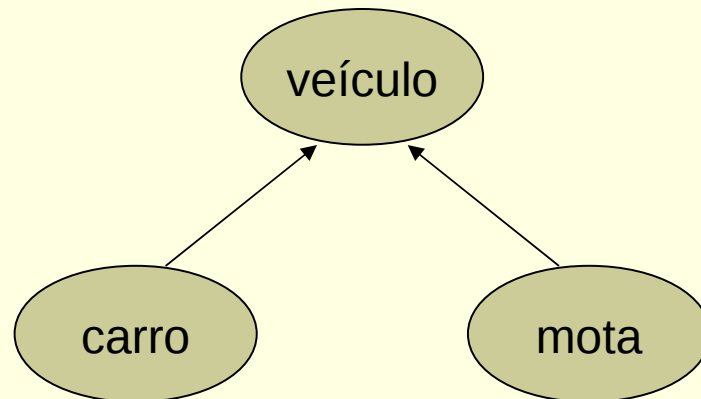
$$P(D) = P'(D) + P(G)$$

IC mede a especificidade de um dado conceito

Teoria de Informação

- Métrica de Resnik

$$\text{sim}(c_1, c_2) = IC(\text{hiper}(c_1, c_2))$$



Teoria de Informação

- Métrica de Lin

$$\text{sim}(c_1, c_2) = \frac{2 \times IC(\text{hiper}(c_1, c_2))}{IC(c_1) + IC(c_2)}$$

- Métrica de Jiang e Conrath

$$\text{dist}_{jcn}(c_1, c_2) = IC(c_1) + IC(c_2) - 2 \times IC(\text{hiper}(c_1, c_2))$$

Teoria de Informação

- IC mede a especificidade de um termo.
 - Então porque não utilizar o número de hipónimos de um termo como medida de especificidade?

$$IC(c) = -\log_2(\text{hypo}(c) + 1)$$

Dicionários

- Utiliza as definições dos dicionários
 - Algoritmo de Lesk
 - Intercessão dos termos contidos nas definições reflecte a associação dos mesmos.
 - **Banco** – “instituição financeira que realiza operações mercantis relacionados com o dinheiro ou com os **títulos** e valores que o representam”
 - **Cheque** – “**título** de crédito que enuncia uma ordem de pagamento da soma nele inscrita”
- Banjeree and Pedersen
 - Utilizam as definições na vizinhança de cada termo no WordNet para desambiguar.

Estudo Comparativo

- Averiguar a semelhança entre pares de palavras.

car	automobile		lad	brother
gem	jewel		journey	car
journey	voyage		oracle	monk
boy	lad		cemetery	woodland
coast	shore		food	rooster
asylum	madhouse		coast	hill
magician	wizard		forest	graveyard
midday	noon		shore	woodland
furnace	stove		monk	slave
food	fruit		coast	forest
bird	cock		lad	wizard
bird	crane		chord	smile
tool	implement		glass	magician
brother	monk		noon	string
crane	implement		rooster	voyage

Estudo Comparativo

Algoritmo	Correlação
Leacock and Chodorow	0,82
Hirst St. Onge	0,68
Banjeree and Pedersen	0,37
Wu and Palmer	0,74
LSA	0,72
Resnik	0,77
Lin	0,80
Jiang and Conrath	-0,81
Resnik*	0,77
Lin*	0,81
Jiang and Conrath*	0,84

WordNet: Relações Semânticas e Métricas de Associação/Semelhança

Seminário Doutoral
Nuno Seco